

MODEL SELECTION UNDER FIXED-BUDGET MISSINGNESS: ORACLE ENVELOPES AND SHARED OBSERVABLE-POLICY AUDITS

Armaan Gupta
Independent Researcher

Abstract : A missingness stress test may allocate unavailable measurements using true outcomes, although a collection process must act before those outcomes are known. We study how this information mismatch affects model selection at fixed measurement counts. Established paired-loss, expected-shortfall, and conditional-allocation results give oracle and information-restricted risk envelopes. An exact binary example shows that oracle-based selection can discard a useful feature and incur 0.0964 nats of excess observable-policy risk. An initial six-dataset study motivates a separately specified extension on six additional public classification datasets. The extension separates predictor fitting, policy learning, selection, and testing; evaluates 14 fixed candidate predictors against the same 61 policy conditions; and retains five splits, two feature blocks, three missingness rates, and four nominal-loss tolerances. At the prespecified half-missingness rate and 0.03-nat tolerance, shared-policy selection has 0.0015 nats lower dataset-averaged held-out worst-library loss than oracle selection, with lower dataset means on 3 of six tasks. Its difference from ordinary nominal-loss selection is -0.0004 nats, and the average advantage over oracle selection reverses when Optical Digits is omitted. These results concern a learned finite policy library, not the optimum over every observable mechanism or a validated real collection system. The contribution is a controlled model-selection assessment and reproducible audit, with explicit separation of known mathematical ingredients from new empirical evidence.

IndexTerms: missingness shift; robust prediction; tabular learning; feature acquisition; conditional value-at-risk; evaluation methodology.

1. INTRODUCTION

A prediction system may receive an optional measurement panel for only a subset of its inputs. A change in operational policy can redirect that panel toward different examples while leaving its overall availability unchanged. For example, an inspection process might continue to measure half its products but change which products qualify for the additional inspection. The model then encounters a different relationship between measurements, missingness, and outcomes. Monitoring the fraction of missing values alone cannot describe this change.

This problem belongs to an established literature. Missingness-shift theory distinguishes conditions under which the optimal predictor remains invariant from conditions under which it changes [1]. StableMiss addresses agnostic mask-distribution shifts through its predictive parameterization and decorrelation [2]. MIRRAMS uses masking and information-motivated loss terms [3], and Knockout supplies a general masking strategy for missing inputs [4]. Adversarial feature deletion predates these methods [5]. These methods motivate testing the marginal value of specialized training against simple controls.

We examine a specific decision question: does selecting a predictor against a label-informed oracle envelope lead to a different and potentially worse choice than selecting against policies restricted to available inputs? A secondary question asks whether the latter assessment adds value beyond ordinary independent-assignment loss. This is more demanding than demonstrating that an oracle risk exceeds an observed-policy risk. It requires a common policy class across candidates, selection without test outcomes, and evaluation on data that did not suggest the hypothesis.

The distinction matters because an oracle can target examples after seeing their outcomes, while an operational rule must act with less information. Conditional feature-value estimation and information-restricted decisions are established in active feature acquisition [25,26] and constrained deferral [27]. Our question concerns their consequence for fixed-count robustness evaluation and model choice. We do not claim to invent these ideas. Nor does a low sensitivity number alone establish useful prediction: ignoring a measurement can reduce sensitivity, while constant positive removal loss can yield zero sensitivity even when the measurement helps.

The study has two stages. Study I supplies the paired-loss audit and an initial six-dataset comparison of robust training with simple controls. Its results motivate the selection hypothesis. Study II specifies that hypothesis before fitting six additional datasets and tests it with four disjoint development/evaluation partitions and a common library of observable policy rules. An exact population example establishes that the model-selection mismatch can occur; the empirical study measures the behavior of specified selectors, including failures and an ordinary nominal-loss comparator. The contribution is this controlled assessment and executable procedure, rather than a new general allocation theorem or a claim of universally superior training.

2. RELATED WORK AND SCOPE OF THE GAP

Missingness shifts. Rockenschaub et al. [1] analyze when a shift leaves the Bayes predictor invariant and emphasize that invariance of the predictor need not imply invariance of its risk. Zhou et al. [14] study domain adaptation under missingness shift. StableMiss [2] treats agnostic changes in mask distributions. These settings establish the importance of acquisition mechanisms. Our audit specializes to a fixed complete-data pool and fixed full mask histogram, so variation in the association between them can be studied without a change in the measurement count. It does not hold the observed-value distribution or the Bayes risk fixed.

Masking and robust prediction. The feature-deletion minimax formulation of Globerson and Roweis [5] establishes a direct predecessor for adversarial removal. Knockout [4] uses placeholder-based masking to learn with missing inputs. MIRRAMS [3] combines original-input classification, classification after extra masking, and confidence-filtered pseudo-label consistency. Our independent-mask baseline follows the established structured-masking principle, while our secondary neural experiment adapts

MIRRAMS's loss ingredients to a common multilayer perceptron (MLP). We do not present either as a new masking method or as a literal reproduction of the original published architecture.

Evaluation and apparent informativeness. Shades-of-Null [6] studies mixed mechanisms and multiple performance dimensions. Jang et al. [16] audit masking with train-only preprocessing and controlled test overlays. ICYM2I [15] addresses biased estimates of modality value under missingness using explicit assumptions and weighting. Our average utility control is an empirical comparison to an always-only predictor, not an estimate of intrinsic information content. The distinguishing audit quantity is the distribution of within-example loss increments, evaluated with unchanged pattern counts and explicit policy-information restrictions.

Related case-level audits. Decision Reliability Profiles [24] use sorted per-case loss differences to assess targeted removal of evaluation cases from a fixed model comparison. Our intervention retains every evaluation row and changes feature availability within that row. The shared additive sorting structure reinforces the need to locate the contribution in the audit setting and interpretation, rather than in sorting itself.

Active feature acquisition and constrained decisions. Gadgil et al. [25] regress incremental prediction-loss improvements on available inputs; their Lemma 1 identifies conditional mutual information under a Bayes predictor and log loss, and Appendix B analyzes predictor error. Our conditional removal-loss regressor uses that established principle for adverse allocation. Valancius et al. [26] condition an acquisition oracle on observed features rather than unrevealed labels and inputs. Charusaie and Samadi [27] give more general constrained randomized decision rules. The threshold rule here is a two-action specialization. Unlike sequential acquisition algorithms, our benchmark holds predictors fixed and tests alternative selection criteria against a common collection-policy library.

Uncertain dependence with fixed marginals. Gao and Kleywegt [28] describe optimization over dependence structures with prescribed marginals and restrictions on their deviation from a reference. Our finite row-by-pattern coupling uses the unrestricted version with fixed pattern frequencies. It does not introduce fixed-marginal robust optimization. The full-information envelope remains a valid conservative target when that unrestricted uncertainty is intended; the empirical question is whether it is a useful selector when operational information is restricted.

Optimization and allocation. The upper-tail representation uses the established conditional value-at-risk (CVaR) formulation [7]. Its integrated summary is a scaled Gini mean difference, consistent with earlier inequality and integrated-risk relationships [8,9]. Conditional score thresholding is closely related to budget-constrained policy learning [10,11]. Multi-pattern assignment uses standard network-flow structure [12], and smoothing uses standard convex composition [13]. We apply these tools to missing-input evaluation and test their practical implications. We make no claim that the mathematical ingredients are novel in isolation. The specific question left to our experiments is the consequence of these choices for held-out model selection under the same finite policy library. We investigate that question without asserting field-wide priority for the combination.

3. MATHEMATICAL ANALYSIS

3.1. Paired losses and the fixed-budget envelope

Let a fixed predictor receive always-observed features Z and an optional feature block S . Write L_0 for its loss when S is observed and L_1 for its loss when S is missing; both losses include the true outcome Y . Define the loss increment $D = L_1 - L_0$. The label Y is the true class, a loss is the penalty assigned to a prediction, and risk is its average. A mask specifies which inputs are hidden. The complete-data population is held fixed. A collection policy chooses a missingness indicator M with $E[M] = q$, so its risk is $E[L_0] + E[MD]$. The model is not retrained when the collection policy changes. Losses are assumed integrable. Finite-sample results require only finite evaluated losses and do not assume independent rows.

For a test pool of n rows and exactly k missing blocks, let $D_{(1)} \leq \dots \leq D_{(n)}$ be the ordered increments and $q=k/n$. The independent-assignment reference is the expected loss over uniformly random assignments with this exact count, not the loss of a particular random assignment.

$$R_{ind}(q) = \overline{L_0} + q\overline{D}. \quad (1)$$

Proposition 1 (sharp empirical envelope). The minimum and maximum loss over all assignments preserving k are

$$R_{min}(q) = \overline{L_0} + \frac{1}{n} \sum_{j=1}^k D_{(j)}, \quad (2)$$

$$R_{max}(q) = \overline{L_0} + \frac{1}{n} \sum_{j=n-k+1}^n D_{(j)}. \quad (3)$$

Proof. For an assignment vector m , the loss is $\text{mean}(L_0) + (1/n)$ times the sum of $m_i D_i$ over rows i . If a maximizing assignment hides a smaller increment and observes a larger one, exchanging their mask labels increases or preserves the objective. Repeated exchanges yield the top- k assignment. Reversing the order gives the minimum. Ties do not affect the achieved value. Empty sums handle $k=0$. This elementary rearrangement argument is used here to obtain the exact endpoints of the risk envelope.

For a numerical illustration, suppose four examples have full-input losses all equal to 0.2 nats and loss increments (0, 0.1, 0.4, 0.9). With two blocks missing, independent-assignment risk is 0.375, minimum risk is 0.225, and maximum risk is 0.525 nats. Hiding the two largest increments produces the maximum. Every assignment still hides exactly two blocks.

The upper premium $V(q)=R_{max}(q)-R_{ind}(q)$ is nonnegative. It measures sensitivity to the association between rows and masks after accounting for average feature-removal loss. It is distinct from $\text{mean}(D)$, which measures the average loss change when the optional block is removed for everyone.

3.2. A rate-integrated sensitivity summary

Extend V linearly between its empirical grid points k/n . This corresponds to fractional assignment probabilities at noninteger budgets. Its slopes decrease, so V is concave with $V(0)=V(1)=0$.

Proposition 2 (integrated premium). If D and D' are independent draws from the empirical distribution of increments, then

$$\int_0^1 V(q) dq = \frac{1}{4} \mathbb{E}|D - D'| = \frac{1}{4n^2} \sum_{i,j} |D_i - D_j|. \quad (4)$$

Proof. Let Q be the empirical quantile function. The top- q term is integral from $1-q$ to 1 of $Q(u)du$. Interchanging the two integrals yields integral from 0 to 1 of $uQ(u)du$. The integral of q mean(D) is mean(D)/2. Therefore the area equals integral $(u-1/2)Q(u)du$. For independent uniforms U, V , integrating $|Q(U)-Q(V)|$ over the two triangles of the unit square gives 2 integral $(2u-1)Q(u)du$, which is four times that area. The identity remains valid for integrable population increments. The Gini mean difference (GMD) is the expected absolute difference between two independent draws. Unlike a normalized Gini coefficient, it retains the units of loss and is well defined even when increments are negative. The displayed expression is the empirical-distribution identity, including diagonal pairs; it is not an unbiased population estimator. This specializes established integrated-quantile/Gini relationships [8,9].

All-rate zero premium is equivalent to constant D , not to ignoring the optional block. In particular, let Y be balanced binary, let the optional feature equal Y , and let Z contain no information. A predictor assigning probability $1-\epsilon$ to the correct class when the feature is present and $.5$ when it is absent has constant $D=\log(2)+\log(1-\epsilon)$. Its premium is zero at every budget, yet it improves on the always-only predictor by $(1-q)D$ for every $q<1$ and $0<\epsilon<.5$. Therefore neither zero premium nor equal risk across mask assignments establishes feature discard. Explicit always-only controls are required.

3.3. Information available to the collection policy

An oracle assignment that sorts the realized D uses outcomes and, in general, optional values. It is a valid stress envelope but may not describe a collection decision that must precede observation of S and Y . Let G be the information available when that decision is made, formally the sigma-algebra generated by the permitted observations. For the observable policy in our main experiment, G contains only the always-observed feature vector. A sigma-algebra specifies which events can be distinguished from those observations. A randomized admissible policy is a G -measurable function taking values in $[0,1]$, with $\mathbb{E}[a]=q$; its Bernoulli randomization is independent of unobserved variables conditional on G . Put $g=\mathbb{E}[D|G]$, the expected loss increment conditional on that information. Write $R_{\max}(q)$ for the population oracle optimum with access to the complete example and label; it equals $\mathbb{E}[L_0]$ plus the integral of the upper q fraction of the quantile function of D . $R_{\text{ind}}(q)=\mathbb{E}[L_0]+q\mathbb{E}[D]$ is the population independent reference.

Proposition 3 (information-restricted envelope). The worst attainable population risk under these policies is

$$R_G(q) = \mathbb{E}L_0 + \inf_t \{qt + \mathbb{E}(g - t)_+\}. \quad (5)$$

Here $(x)_+=\max(x,0)$, and an upper-tail mean averages the largest specified fraction of a distribution. For $0<q<1$, the added term is q times the upper-tail mean of g , with fractional weighting of quantile atoms. A threshold rule on g attains the optimum, with randomized assignment at the threshold when necessary. If G is contained in H , then $R_{\text{ind}}(q) \leq R_G(q) \leq R_H(q) \leq R_{\max}(q)$. In particular, the integrated oracle-minus-observable gap is one-quarter of $\text{GMD}(D)-\text{GMD}(g)$, and is nonnegative.

Proof. The tower property gives $\mathbb{E}[aD]=\mathbb{E}[ag]$. For every t , $ag \leq (g-t)_+ + at$, giving the displayed upper bound. A top- q threshold policy with randomized ties attains equality. Inclusion of policy classes proves the ordering. Alternatively, conditional Jensen gives $\mathbb{E}[(\mathbb{E}[D|G]-t)_+] \leq \mathbb{E}[(D-t)_+]$. Apply Proposition 2 to g and D for the integrated result. This is the standard expected-shortfall variational identity [7] applied to a conditional loss increment. Conditional threshold allocation is also central to budget-constrained policy learning [10,11]. No causal treatment-effect identification is required here because both masked model losses can be evaluated on complete development examples; masking does not change their recorded labels.

The empirical oracle and a learned observable policy are not directly interchangeable with the two population optima. The learned policy can be suboptimal; therefore their observed gap combines restrictions on information, estimation error, and finite-pool effects. Our experiments report this as an oracle-versus-learned-policy gap, not as an identified population value of information.

Corollary (plug-in allocation regret). Suppose $\hat{a}$ maximizes $\mathbb{E}[a g\text{-hat}]$ over the same budget-constrained policy class and $\hat{a}\text{-star}$ maximizes $\mathbb{E}[a g]$. Then

$$0 \leq \mathbb{E}[(\hat{a}^* - \hat{a})g] \leq \mathbb{E}|g - \hat{g}|. \quad (6)$$

Proof. Decompose the regret as

$$\mathbb{E}[(\hat{a}^* - \hat{a})g] = \mathbb{E}[(\hat{a}^* - \hat{a})(g - \hat{g})] + \mathbb{E}[(\hat{a}^* - \hat{a})\hat{g}]. \quad (7)$$

The second term is nonpositive by optimality of $\hat{a}$; the first is at most $\mathbb{E}|g-\hat{g}|$ because both policies lie in $[0,1]$. An infinity-norm alternative follows from $\mathbb{E}|\hat{a}\text{-star}-\hat{a}| \leq 2 \min(q,1-q)$: the regret is at most $2 \min(q,1-q)$ times the largest absolute score error. This deterministic comparison assumes both policies meet the same population budget. Test-pool ranking enforces an empirical budget instead and does not automatically inherit that population guarantee.

3.4. More than two availability patterns

For K patterns, form the n -by- K loss matrix L . If pattern j must occur k_j times, solve the transportation program

$$\max_{A \geq 0} \frac{1}{n} \sum_{i,j} A_{ij} L_{ij}, \quad A1 = 1, \quad A^T 1 = k. \quad (8)$$

Here A_{ij} is the assignment weight of row i to pattern j ; the vector 1 contains ones, and the entries of k are the prescribed pattern counts. The constraints give each row total weight one and preserve every pattern count. Integer row supplies and pattern demands give an integral optimal solution through the standard transportation polytope; a linear-programming solver can recover a valid assignment [12]. The independent reference is the weighted sum of column means with weights k_j/n . Feasibility restrictions can

remove prohibited row-pattern pairs, but our experiments assume every pattern is feasible because development and auditing data are complete.

Proposition 4 (assignment invariance). For the finite fractional transport polytope with strictly positive row and pattern marginals, every feasible assignment has the same objective if and only if $L_{ij} = u_i + v_j$.

Proof. Additive losses have constant objective because both marginals are fixed. Conversely, the product coupling is an interior point. A sufficiently small positive or negative 2-by-2 cycle perturbation preserves the marginals. Invariance forces the cycle identity

$$L_{ij} + L_{i'j'} - L_{ij'} - L_{i'j} = 0. \quad (9)$$

This holds for every pair of rows and columns. Fixing one row and one column yields the additive representation. This is a standard transportation-cycle characterization. For two patterns it reduces to a constant loss increment. Structural-zero feasibility constraints require a modified cycle condition and are outside this statement.

3.5. Smoothed robust training

For a linear-softmax model with parameters w , define the exact empirical oracle risk at q by the envelope above. The expected-shortfall variational form gives

$$R_{\max}(w, q) = \min_t \{qt + \frac{1}{n} \sum_i \max\{L_{i0}(w), L_{i1}(w) - t\}\}. \quad (10)$$

For $0 < q < 1$ and $\tau > 0$, replace each maximum with temperature- τ log-sum-exp. Denote the resulting threshold objective by $R_{\tau}(w, t, q)$, and its minimum over t by $R_{\tau}(w, q)$. Jointly optimize w and t for

$$J(w, t) = (1 - \alpha)R_{\text{ind}}(w, q) + \alpha R_{\tau}(w, t, q) + \frac{\lambda}{2} \|w_{\text{pen}}\|_F^2. \quad (11)$$

Here α is the robust-loss weight, λ is the nonnegative L2 penalty strength, τ is the smoothing temperature, and w_{pen} contains all coefficients except intercepts. The Frobenius norm sums the squared entries before taking a square root. For nonnegative $\alpha \leq 1$ and convex per-pattern losses, J is jointly convex: log-sum-exp is convex and coordinatewise increasing, and each argument is a convex loss plus an affine threshold. For scalar arguments a and b , $\max(a, b) \leq \tau \log(\exp(a/\tau) + \exp(b/\tau)) \leq \max(a, b) + \tau \log(2)$. Applying this inequality row by row and minimizing over t gives $R_{\max} \leq R_{\tau} \leq R_{\max} + \tau \log(2)$. Thus the smoothing error in the mixed objective is at most $\alpha \tau \log(2)$. These statements do not assert finite-sample generalization, nor that an optimizer stopped at a numerical tolerance has reached an exact minimum. We log convergence and gradients and evaluate the unsmoothed test envelope. The smoothing and convex-optimization tools are established [7,13]; the method is a transparent comparator within our audit, not a claimed new state of the art.

3.6. Audit procedure

Inputs. A fixed predictor, complete held-out examples and labels, two prespecified availability states, a missing-block count k , and the information permitted to the collection policy.

Paired evaluation. Evaluate each row under both states. Form its loss increment and retain its original row identifier. Compute independent risk, the sorted minimum and maximum, and the integrated premium. These quantities require two prediction evaluations per row and $O(n \log n)$ sorting time; retaining the increments uses $O(n)$ additional storage.

Policy restriction. Using development data only, estimate loss increments from the permitted pre-measurement inputs. Rank test rows by these scores and assign exactly k missing blocks. Report the achieved loss separately from the label-informed oracle.

Utility control and reporting. Report the full-input and missing-input losses, independent risk, oracle premium, learned-policy risk, and a separately tuned always-only control. Prespecify the budget and model-selection rule, and retain paired per-row results. This procedure audits a fixed model on a complete pool; it does not infer unobserved counterfactual features from naturally incomplete records.

3.7. A valid envelope can select the wrong model for a different policy class

Risk-envelope validity does not imply that minimizing the envelope selects the best model for an information-restricted objective. Consider balanced binary Y , an independent observed bit Z , and an optional bit S obtained by independently flipping Y with probability 0.2. Candidate A outputs the true conditional class probabilities given S when it is present and probabilities (0.5, 0.5) otherwise. Candidate B always outputs (0.5, 0.5). Let $h(p) = -p \log(p) - (1-p) \log(1-p)$, and let $q = 0.5$.

For A, full-input risk is $h(0.8) = 0.5004$ nats. The removal increment is $\log(1.6)$ on the 80% of examples where $S=Y$ and $\log(0.4)$ on the remaining examples. Every policy observing only Z has the same expected increment because Z is independent of (S, Y) . Thus its attainable risk is the independent risk, $(h(0.8) + \log(2))/2 = 0.5968$. The label-informed oracle instead assigns missingness to half the population within the first group and attains $h(0.8) + 0.5 \log(1.6) = 0.7354$. B has risk $\log(2) = 0.6931$ under every assignment. Oracle minimization over these two candidates therefore selects B, although A has lower risk under every admissible Z -based policy. The observable excess from that choice is $(\log(2) - h(0.8))/2 = 0.0964$ nats.

This counterexample separates objectives; it does not show that oracle minimization is incorrect for its own stated uncertainty set. It is a direct application of the conditional-allocation framework, not a new mutual-information identity. The appendix retains the prespecified accuracy and budget grid and a heterogeneous two-group extension, so the single example is not treated as a frequency estimate.

For a finite library P of already-fitted score policies, define $W_P(f)$ as the largest evaluated risk over policies in P . The shared-policy selector minimizes its selection-set estimate over the same validation-eligible candidate set used by the comparator selectors. Every candidate is evaluated under every library assignment. The maximizing library member may differ by candidate: a common uncertainty class does not mean one universally worst physical assignment. W_P is an achieved-library envelope, not a certified

optimum over all observable policies. Exact-count ranking acts on the observed evaluation batch and its tie-breaking order; it is not automatically a population-threshold policy covered by Proposition 3.

4. EXPERIMENTAL METHODOLOGY

4.1. Study I: initial questions and protocol

The primary analysis asks whether independent masking leaves a substantial fixed-budget loss envelope, whether a learned policy using always-observed features attains the oracle envelope, and whether smoothed robust linear training improves the validation-selected trade-off beyond blending. Nominal risk denotes expected loss under independent assignment at the same budget; oracle risk denotes the largest loss over all assignments at that budget. Lower risk is better for the predictor. The learned policy deliberately seeks higher loss as a stress test; it is not a proposed beneficial collection rule.

The protocol and main implementation were hashed locally before the benchmark began. This is an internal frozen protocol, not external preregistration. Earlier experiments on Wine Quality and a synthetic binary task informed the design; neither wine subset is included in the new benchmark. Subsequent pattern-specific experts and the heterogeneous finite-distribution simulation are labeled supplementary. The initial plan for a full Transformer comparison was replaced, before neural results, by an MLP objective comparison after a CPU-only runtime probe. All deviations are retained in the reproducibility archive.

4.2. Data and information access

We use six public UCI datasets [17-22], retaining their classification targets and finite numerical features. Table 1 reports their dimensions. MAGIC is generated by a physics simulation. The other datasets supply recorded feature measurements, but none documents the acquisition-policy shifts imposed here. The feature panels are synthetic experimental partitions, not validated independent sensors; this is particularly relevant where several features are derived from the same image.

Table 1. Dataset sizes before partitioning. All tasks retain their original classes. Full split indices, grouped-duplicate checks, excluded cross-partition duplicates, and dataset hashes accompany the results.

Dataset	Rows	Features	Classes	Partition
Banknote	1372	4	2	Grouped 60/20/20
Spambase	4601	57	2	Grouped 60/20/20
MAGIC	19020	10	2	Grouped 60/20/20
Pendigits	10992	16	10	Official test
Dry Bean	13611	16	7	Grouped 60/20/20
Satellite	6435	36	6	Official test

Every method may use complete labeled development data. Model fitting uses only the training partition; hyperparameter selection and learned-policy construction use validation data. At prediction time a model receives the available values and missingness indicators. The evaluator uses complete held-out examples to calculate both potential masked losses, but hidden values and labels are not passed to the masked predictor or to the learned test-policy score. This complete-development assumption is essential: arbitrary remasking cannot be performed on naturally incomplete data without additional identification or reconstruction assumptions.

For Banknote, Spambase, MAGIC, and Dry Bean, we create approximately 60/20/20 group-stratified train/validation/test partitions. Identical full feature vectors share a group even if labels differ; group strata use their modal class. Pendigits and Satellite retain official test partitions; any development row whose complete vector also occurs in the test partition is excluded, and the remaining development data are split 80/20. The Pendigits official test is writer-disjoint from its training set [20]. Satellite's overlapping neighborhoods and unavailable coordinates prevent a claim of spatially independent evaluation [22]. Mean and standard deviation are estimated from training rows only; missing values are represented by standardized zero plus feature indicators. We do not winsorize inputs or tune preprocessing on test data.

The six tasks were selected to cover binary and multiclass numerical classification, a range of feature dimensions, and both constructed and official partitions while permitting a complete local experiment. They are not a representative sample of deployment systems.

There are five seeds and two panels per dataset, totaling 60 dataset/split/panel configurations. Each optional panel contains $\text{floor}(d/2)$ features selected by a fixed label-independent feature permutation. Panels use the same seed across datasets but act on each dataset's feature dimension. Within each configuration the same held-out rows are reused across all policies. This paired design controls the complete-data pool; it does not make the rows, seeds, or repeated policies independent replications.

4.3. Models, tuning, and controls

Linear models are multiclass softmax predictors: each class receives a linear score, and exponentiating and normalizing these scores gives class probabilities. Inputs are zero-filled standardized features, per-feature missingness indicators, and an intercept. An indicator equals one only when its corresponding value has been hidden. Independent two-pattern training evaluates both versions of every training row and weights their losses by $1-q$ and q , where $q=\text{floor}(n/2)/n$ for that partition. Thus the linear and tree baselines integrate this masking expectation rather than relying on a single random training mask. The always-only control removes the optional values and all missingness indicators in both states. We evaluate independent two-pattern training, an always-observed-only model, a model trained on a label-associated source mask, extra deletion from that source mask, and smoothed robust training. Regularization values are 0.0001, 0.001, and 0.01. Robust weights are 0, 0.10, 0.25, 0.50, 0.75, and 1; smoothing temperature is 0.02. The limited-memory BFGS optimizer (L-BFGS) minimizes the convex objective with a maximum of 2,000 iterations, function

tolerance $10^{(-12)}$, gradient tolerance $10^{(-7)}$, and at most 40 line-search steps. Intercepts are not penalized. The exact, unsmoothed envelope is used for model selection and reporting, so smoothing does not redefine the reported adversary.

For each robust weight, regularization is chosen using the corresponding mixture of nominal and exact-worst validation risks. Independent and always-only models are tuned on nominal validation loss. Source and extra-deletion diagnostics are tuned on their respective source objectives. The source mask hides $\text{floor}(n/2)$ rows by ranking a score equal to twice the signed label-group indicator plus logistic noise; multiclass groups use parity of the sorted class code. In the extra-deletion diagnostic, an already-missing block remains missing and an observed block is additionally hidden with probability 0.5; the resulting probabilities are integrated in the training objective. This deliberately injected shortcut is an evaluator-designed stress condition.

The nonlinear comparator is histogram gradient boosting from scikit-learn [23]: 200 iterations, learning rate 0.08, 15 or 31 leaves, L2 regularization 0 or 1, minimum 20 samples per leaf, and no internal early stopping. The four configurations are tuned on nominal validation loss. Independent and always-only tree models are tuned separately. Complete development data and the same masks are available to all methods; the comparison therefore controls access to development information.

For each backbone, we combine its independent and always-only predictions using both arithmetic probability mixtures and logit mixtures, with blend weights $\beta=0, 0.05, \dots, 1$. A probability blend is $(1-\beta)p_{\text{ind}} + \beta p_{\text{always}}$. A logit blend applies softmax after mixing the two class-score vectors with the same weights. For trees, these scores are log probabilities. Predicted tree probabilities are clipped below at $10^{(-12)}$, then renormalized through softmax when computing loss; linear and neural losses use a numerically stable log-sum-exp directly on their class scores. For each nominal-loss tolerance 0, 0.01, 0.03, or 0.10 nats above the independent model's validation risk, we select the candidate with the lowest worst-assignment validation loss among feasible candidates. The independent model belongs to every family, ensuring feasibility. Robust-linear and blend-linear selections are independent and evaluated unchanged on the test set. This matches validation tolerances, not realized test losses. We explicitly report when the tolerance fails on held-out data; a favorable test point from the full grid is never substituted for the validation-selected point.

4.4. Policies and endpoints

The primary loss is cross-entropy, also called log loss: the negative natural logarithm of the probability assigned to the true class. Natural logarithms give units of nats. A correct-class probability of 0.5 incurs about 0.693 nats, whereas 0.9 incurs about 0.105 nats; confident mistakes receive larger penalties. Secondary common-policy outputs include accuracy, the sum-over-classes Brier score (squared probability error), and area under the receiver-operating-characteristic curve (AUROC; binary or macro one-versus-rest). The reassignment formulas optimize additive log loss. They do not make the same assignment worst-case for AUROC, which is not an additive per-row loss. Accuracy is not interpreted as the domain-specific operating objective for MAGIC or Spambase, whose error costs are asymmetric [18,19].

We calculate the exact two-pattern envelopes at missing fractions 0.10, 0.25, 0.50, 0.75, and 0.90, rounding the number of missing blocks down. The paired-loss formula calculates the independent-assignment expectation exactly; a separately sampled random policy is reported as a common-policy check. Other common half-budget policies rank a positively or negatively signed always-observed feature plus logistic noise, or use outcome-associated scores for labeled stress tests. All preserve the same counts. Outcome-based and oracle policies are explicitly separated from feasible pre-measurement policies.

For the learned observable policy, a 100-iteration gradient-boosting regressor estimates the validation loss increment from always-observed features only. It uses seven leaves, a minimum of 30 rows per leaf, L2 penalty 1, and learning rate 0.05. Test rows are ranked by the predicted increment, with label-independent random tie-breaking, and $\text{floor}(n_{\text{test}}/2)$ rows lose their optional block. A separate policy regressor is fitted for each audited predictor because its loss increments differ. Comparisons between learned-policy columns therefore involve model-specific stress policies, not one shared operational assignment. Validation data are reused for predictor selection and policy fitting; only the test partition evaluates achieved policy risk. The trained predictor is fixed. The policy acts on a batch of observed test features to enforce its budget; it is not an online arrival-by-arrival rule. Because the regressor is imperfect, its risk need not reach the population optimum or even exceed the independent reference on every finite test pool.

4.5. Supplementary analyses and verification

The four-pattern extension splits one optional panel into two subpanels and uses all four combinations of availability. For seed 0/panel 0 on each task, independent linear and tree models are fitted to all four patterns and tuned by mean validation loss. Pattern quotas are as equal as possible. Exact transportation programs give the upper and lower envelopes; a separate loss regressor for each pattern supplies the observable-policy scores. The solver receives score matrices rather than test labels for learned-policy assignment.

The neural sensitivity analysis uses a multilayer perceptron (MLP) with two 128-unit hidden layers and rectified linear unit (ReLU) activations. All three objectives use feature indicators and the same initialization: independent block masking; the MIRRAMS three-term objective starting from complete inputs; and the same objective starting from independently block-masked inputs. Additional masks independently remove each input feature with probability 0.2. Both additional loss weights are 15. The pseudo-label is the highest-probability class from the original-input prediction and is used only when that probability is at least 0.95. This teacher prediction is detached from gradient computation. The confidence-filtered term is averaged over the entire batch, with zero contribution from rows below the threshold. Adam uses learning rate 0.0001 and batch size 256. Validation loss is checked every five epochs, with a 1,000-epoch ceiling and 100-epoch patience. This is a single-seed, common-backbone objective adaptation; it is not evidence about the full original Transformer or a reproduction of its reported scores.

Pattern-specific logistic experts, added as a supplementary diagnostic, separately learn each availability pattern from complete training examples and tune their regularization independently. They assess whether sharing linear coefficients across patterns restricts the primary baseline. The finite-distribution simulations evaluate known predictors and policy-specific Bayes risk exactly. Here Bayes risk means the lowest attainable average loss for a predictor that knows the distribution of the observed inputs, mask, and label.

For the heterogeneous simulation, Y and a group variable H are independent balanced binary variables. The always-observed bit B and optional bit S are conditionally independent given (H, Y) . Their probabilities of matching Y are respectively $(0.65, 0.95)$ when $H=0$ and $(0.85, 0.65)$ when $H=1$. The policy observes (H, B) . The fixed predictor uses the exact conditional class probability under independent masking: $P(Y=1|H, B, S)$ when S is present and $P(Y=1|H, B)$ otherwise. All 16 states have known probabilities. At $q=0.5$, exact allocation integrates these probabilities, including fractional assignment for tied score groups. Learned scores are sample means of the paired loss increment within each of the four (H, B) groups, with zero used for an empty group. Samples of 100, 500, 2,000, and 10,000 are drawn under 50 seeds. Allocation uses the known population group masses, isolating score estimation from estimation of the budget threshold.

The second simulation has eight states: Y is balanced; the always-observed bit and optional bit match Y with probabilities 0.7 and 0.9 and are conditionally independent given Y . We hold the independent-mask Bayes predictor fixed while changing the policy and separately recompute each policy-specific Bayes risk. This distinguishes a changed prediction problem from error in a fixed predictor.

Gradient checks use both finite differences and independent automatic differentiation. Small assignment problems are checked against exhaustive enumeration; integrated-premium, smoothing, convexity, conditional-information ordering, and plug-in regret identities receive numerical checks. Prediction files retain every fitted main-grid candidate. A separate audit recomputes summaries and selection rules from those files. The tests check implementation consistency; they do not establish external validity or constitute independent peer review.

4.6. Statistical interpretation

Dataset-level means average the ten seed/panel configurations. Paired method differences and their descriptive standard deviations are supplied by dataset and tolerance. The six datasets are a convenience sample, so cross-dataset averages describe this collection rather than a population of ML tasks. We do not report significance tests that treat masks, overlapping splits, or repeated official test sets as independent observations. In particular, the fixed Pendigits and Satellite tests recur across seeds, while Satellite may contain spatial dependence. Full per-configuration results allow readers to inspect heterogeneity rather than rely on a pooled average.

4.7. Study II: a separately specified common-policy selection test

The new primary comparison is shared-policy selection minus oracle selection in held-out worst-library log loss, at $q=0.5$ and nominal tolerance 0.03 nats. We specified this endpoint, datasets, candidates, seeds, policies, and analysis before the first new predictive fit and saved a timestamped code/protocol hash lock. This was local protocol freezing, not external preregistration. Study I is exploratory for this hypothesis. The six additional datasets were not used to design it or to select the earlier models.

Table 2. New datasets [29-34] and documented feature groups for Study II. WDBC denotes Breast Cancer Wisconsin (Diagnostic). White Wine is binarized at quality at least six. Vehicle's downloaded files contain 846 rows despite a 946-row UCI landing-page count; all downloaded rows are retained. Blocks are interpretable groups of recorded features, not established independently purchased measurements.

Dataset	Rows	Features	Classes	Documented optional block
WDBC	569	30	2	Worst-value summaries (10)
White Wine	4898	11	2	Chloride/sulfur/sulphate (4)
Optical Digits	5620	64	10	Right image half (32)
Page Blocks	5473	10	5	Pixel/run descriptors (6)
Vehicle	846	18	4	Higher moments (7)
Steel Faults	1941	27	7	Luminosity descriptors (4)

Each dataset uses split seeds 410-414 and two panels: the documented block and a same-sized random block generated with seed 6281. Identical full feature vectors, regardless of label, stay in the same partition. Nonofficial data are split approximately 50% for predictor fitting, 20% for policy fitting, 15% for model selection, and 15% for testing, using modal-label stratification of duplicate groups. Optical Digits retains its official 1,797-row test; its 3,823 training rows are split approximately 60/20/20. Any development duplicate of an official test row is excluded. Only predictor-fitting rows determine feature standardization. The archive records all indices and actual counts; small differences from target proportions reflect groups and integer rounding.

The same 14-candidate set is used for every selector. Four endpoints are independently masked and always-only linear softmax and gradient-boosting predictors. Linear L2 is 0.001; boosting has 200 iterations, learning rate 0.08, 31 leaves, L2 1, minimum leaf size 20, and no early stopping. Four additional linear candidates use the smoothed robust objective at α 0.25, 0.5, 0.75, and 1, with L2 0.001 and temperature 0.02. Six candidates mix probabilities from the independent and always-only endpoints, separately by backbone, at always-only weights 0.25, 0.5, and 0.75. Predictor fitting uses only the first partition. The training rate remains one-half at every evaluation budget. No hyperparameter is chosen using test results.

On the second partition, each candidate supplies paired removal losses. Each of the 14 increment targets is regressed on always-observed features using both histogram gradient boosting (HGB; 100 iterations, seven leaves, minimum leaf size 30, L2 1, learning rate 0.05, no early stopping) and ridge regression (penalty 10). Each score and its negative define two rank policies, yielding 56 policies. Four further policies rank by the first always-observed feature or the first principal component, in either direction; the component is fitted on predictor-training data. Uniform exact-count assignment, evaluated in expectation, gives the 61st condition. These are established regression/allocation constructions [25,27], not a proposed new acquisition algorithm.

All candidates face this same library. Each score policy hides $\text{floor}(q \cdot n)$ blocks in a batch of n examples. It observes only always-available features and uses row order to break ties; fitted score functions are frozen before selection. Test labels and optional test

inputs are not policy inputs. Evaluation includes $q=0.25, 0.5$, and 0.75 . The smaller boosting-only and ridge-only libraries each contain their 28 learned score policies, the four feature policies, and uniform assignment. They provide estimator ablations and tests against an alternative score family.

On the third partition, a candidate is eligible when its independent-assignment loss is within tolerance of the lowest such candidate loss. All selectors use this same eligible set, at tolerances 0, 0.01, 0.03, and 0.10 nats. They minimize respectively nominal loss, the full-information oracle envelope, the maximum over the shared library, the boosting-only maximum, or the ridge-only maximum. Candidate order resolves objective ties. Selections are written before final test evaluation. Thus model-specific policy fitting supplies members of a common library rather than a different single achieved-policy target for each model.

Held-out outcomes are nominal risk, oracle risk, and maximum risk over each fixed library. Retrospective selection regret subtracts the best test worst-library risk among the validation-eligible candidates; this finite-pool hindsight comparator is descriptive and unavailable at selection time. A test nominal violation compares the selected model with the best test nominal candidate at the same tolerance. Neither quantity supplies a population guarantee. Dataset means average ten overlapping configurations, then receive equal weight. All rate/tolerance results, configuration wins/ties/losses, and leave-one-dataset-out means are retained. These are descriptive estimates on six tasks, without treating sixty configurations as independent studies.

A secondary source-order sensitivity check replaces deterministic tie order with ten fixed pseudorandom permutations at the primary budget/tolerance. It repeats selection and evaluation using the saved scores and predictions, without refitting or replacing the primary results. This diagnostic was specified while fitting was underway, before outcome summaries were examined, and is recorded separately from the frozen primary protocol. It matters because source ordering can convey structure even when a score is constant.

5. RESULTS

5.1. Study I: fixed-count assignment sensitivity

Table 3. Half-budget test log loss, averaged over ten configurations per dataset. The final column is an empirical nominal-loss benefit relative to a separately fitted always-only model; positive values favor optional-feature use. It is not an intrinsic information estimate.

Dataset	Model	Independent	Oracle	Learned policy	Always-only minus independent
Banknote	linear	0.1901	0.3435	0.3045	0.1218
Banknote	tree	0.1418	0.2573	0.2351	0.1321
Spambase	linear	0.3305	0.4407	0.3523	0.0367
Spambase	tree	0.1722	0.2335	0.1976	0.0328
MAGIC	linear	0.4596	0.4999	0.4710	0.0039
MAGIC	tree	0.3260	0.3973	0.3390	0.0263
Pendigits	linear	0.5244	0.7230	0.6122	0.1111
Pendigits	tree	0.2307	0.2836	0.2530	0.0467
Dry Bean	linear	0.2266	0.2423	0.2295	0.0010
Dry Bean	tree	0.2380	0.2487	0.2387	0.0027
Satellite	linear	0.3896	0.4434	0.3982	0.0041
Satellite	tree	0.3036	0.3227	0.3057	0.0054

Independent masking retained a positive mean oracle premium for every dataset and backbone. Linear premiums ranged from 0.0157 nats on Dry Bean to 0.1986 on Pendigits; tree premiums ranged from 0.0107 on Dry Bean to 0.1155 on Banknote. The learned observable-policy risks were below the oracle risks in every dataset/backbone mean. These differences cannot be attributed solely to unavailable information because the learned regressor may also be imperfect. Optional-feature benefit was heterogeneous: the linear nominal advantage over the always-only control was 0.1218 nats on Banknote but only 0.0010 on Dry Bean.

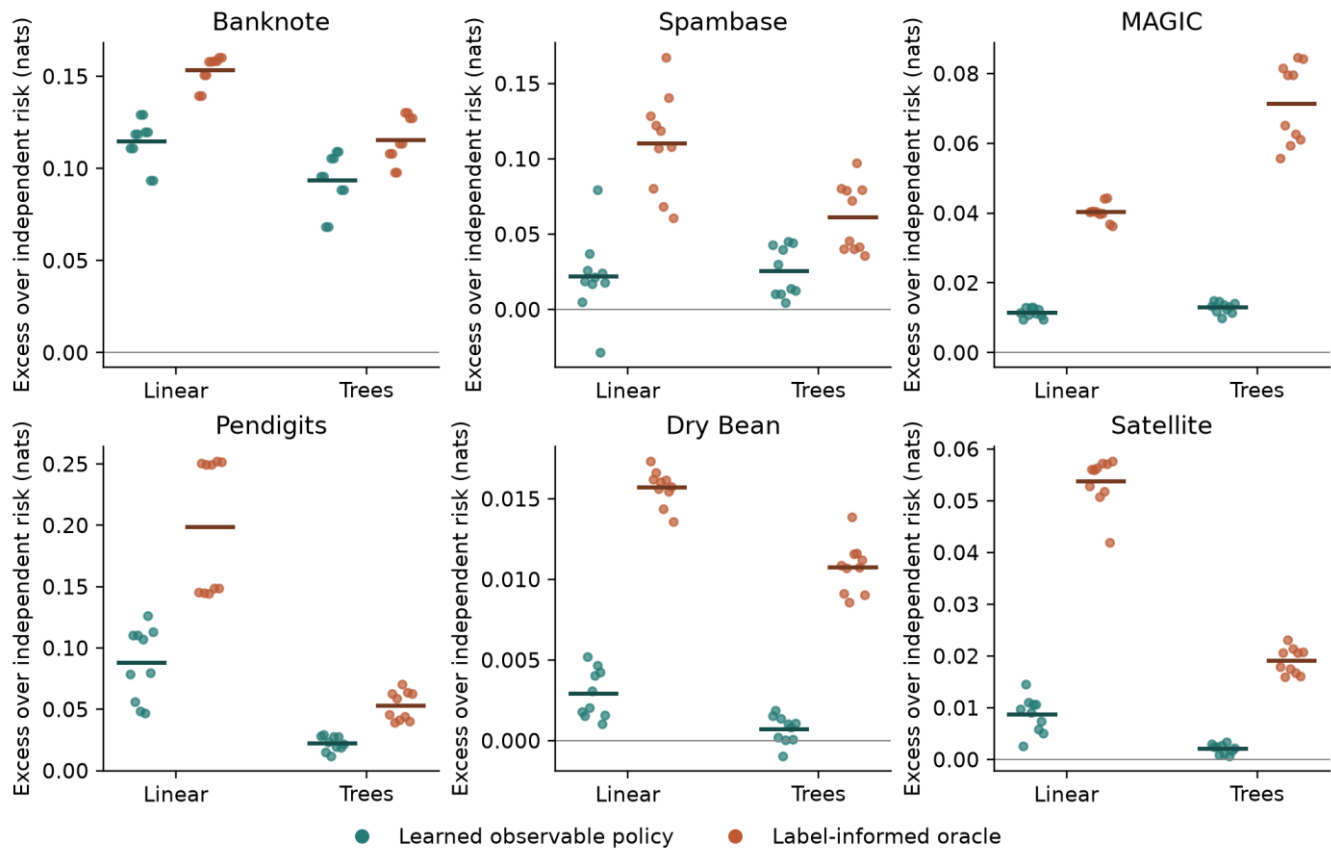


Figure 1. Each point is one split/panel configuration. Green points show achieved learned-policy increments and orange points show oracle premiums over independent assignment. Short horizontal bars indicate configuration means for each backbone.

5.2. Dedicated robust training does not dominate blending

Table 4. Robust linear minus blended linear test risk, first averaged within dataset and then equally across six datasets. Negative differences favor robust training. Wins use a numerical threshold of 0.000001 nats. Test violations mean the realized nominal-loss increase exceeds the validation tolerance by more than that threshold.

Validation tolerance	Mean oracle difference	Mean nominal difference	Dataset oracle wins	Configuration wins	Robust test violations
0.0000	0.0051	0.0004	2/6	13/60	10/60
0.0100	0.0029	-0.0021	0/6	8/60	5/60
0.0300	0.0020	-0.0042	0/6	10/60	1/60
0.1000	0.0030	-0.0105	0/6	4/60	1/60

At all three positive tolerances, blended linear models had lower mean oracle loss on every dataset. At tolerance 0.03, robust training had 0.0020 nats higher oracle loss and 0.0042 nats lower nominal loss on the equally weighted dataset average. It won the oracle comparison in 10 of 60 paired configurations. Thus the results show different trade-offs, not strict dominance of either procedure at matched realized test loss. At zero tolerance robust training won two dataset means, but its pooled oracle difference remained positive.

The blend family has 40 candidates after tuning the two endpoint models, whereas the robust family has six regularization-selected weights. Their search capacities and fitting budgets differ; this is a comparison of specified procedures, not a matched-compute or exhaustive-frontier theorem. The 0.03 row is discussed for illustration; all four protocol tolerances are reported, without post-test replacement of selected models.

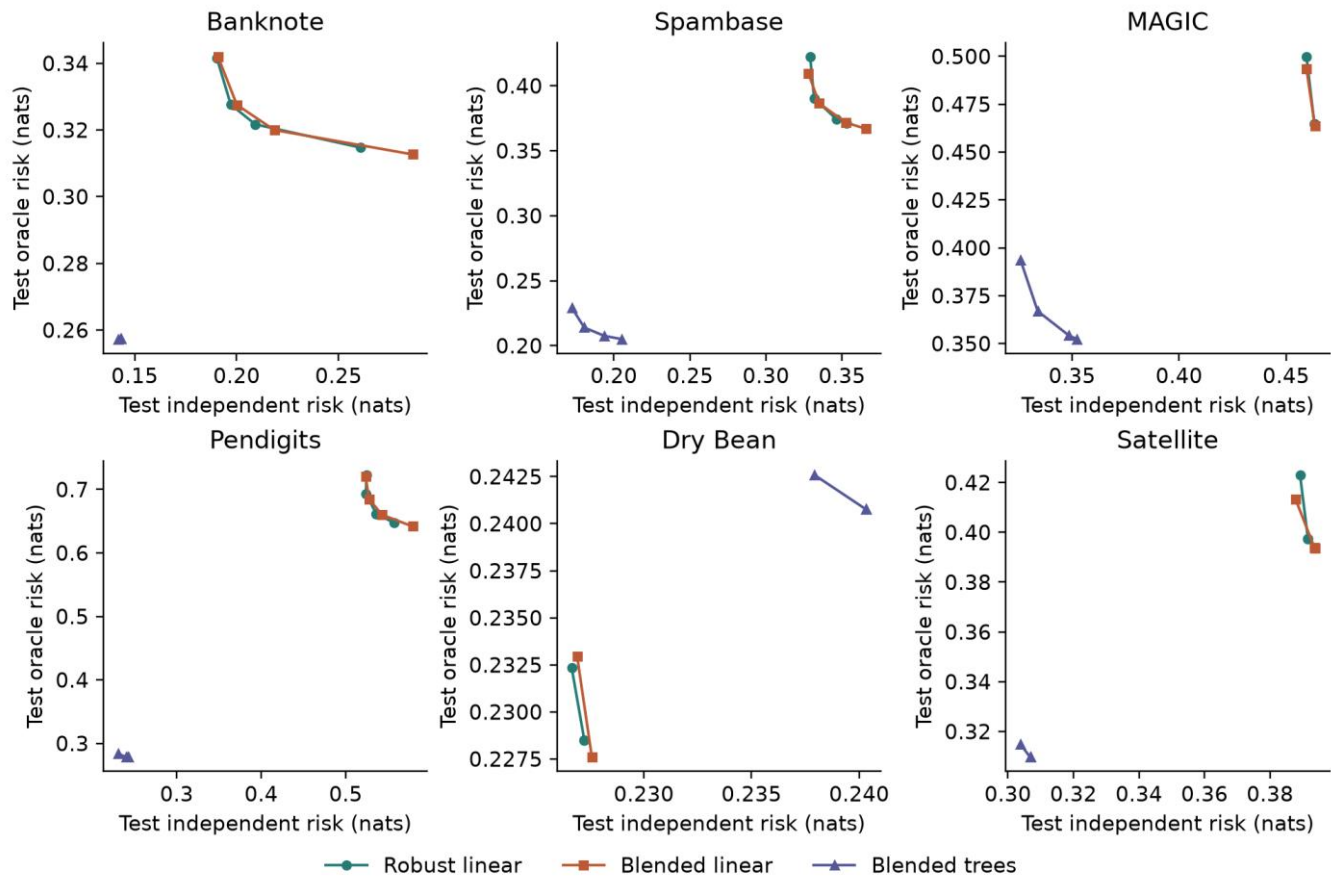


Figure 2. Dataset-mean test results of validation-selected candidates at the four tolerances. Connecting points describes the selection procedures; the curves are not optimized test frontiers. Tree blends provide an additional nonlinear reference.

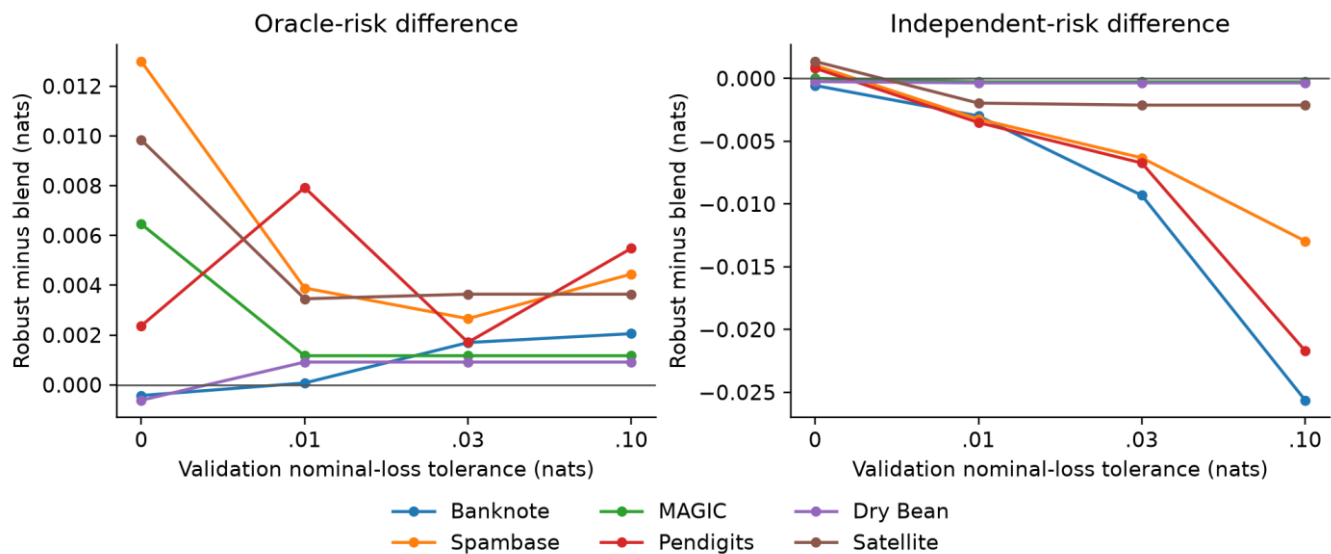


Figure 3. Dataset-level paired robust-minus-blend differences at every tolerance. The supplementary tables retain configuration variability.

5.3. Supplementary controls

Table 5. Mean test log losses in nats for separately trained pattern experts compared with the primary shared-coefficient linear baseline; ten configurations per dataset. This supplementary comparison was added after the main experiment began. Differences reflect changed model restrictions and separate regularization selection, so they should not be interpreted as an acquisition-information limit.

Dataset	Experts nominal	Shared linear nominal	Experts oracle	Shared linear oracle
Banknote	0.1727	0.1901	0.3102	0.3435
Spambase	0.3203	0.3305	0.4571	0.4407
MAGIC	0.4592	0.4596	0.5009	0.4999
Pendigits	0.4957	0.5244	0.7215	0.7230

Dataset	Experts nominal	Shared linear nominal	Experts oracle	Shared linear oracle
Dry Bean	0.2236	0.2266	0.2394	0.2423
Satellite	0.3869	0.3896	0.4480	0.4434

The four-pattern experiments also produced nonzero assignment envelopes on every dataset and backbone. Their learned score-based transport assignments stayed within the exact oracle bounds. This establishes implementation feasibility for four patterns on these datasets, not scalability to arbitrarily many patterns.

Table 6. Oracle test log loss (nats; lower is better) in the single-seed common-MLP sensitivity analysis. The latter two objectives adapt MIRRAMS loss ingredients; they are not the original Transformer implementation.

Dataset	Independent MLP	Complete-start objective	Block-start objective
Banknote	0.2321	0.2858	0.2586
Spambase	0.3035	0.3219	0.3025
MAGIC	0.4206	0.4936	0.4259
Pendigits	0.2311	0.3505	0.2460
Dry Bean	0.2121	0.2961	0.2278
Satellite	0.3448	0.4713	0.3494

The complete-start adaptation had higher oracle loss than independent masking on all six tasks. The block-start adaptation had lower oracle loss only on Spambase. The objective adaptation and one-seed design limit these findings to the reported configurations. Of 18 neural runs, 10 reached the 1,000-epoch ceiling; each reported checkpoint was selected by validation nominal loss.

5.4. Exact information separation and verification

In the heterogeneous finite-distribution simulation, independent risk was 0.414775, optimal observable-policy risk was 0.523335, and oracle risk was 0.592591 nats. The exact oracle-minus-observable gap was therefore 0.069256 nats. Across 50 seeds at each training size 100, 500, 2,000, and 10,000, learned group-mean scores all recovered the correct allocation ranking and attained the observable optimum. The flat learning curve indicates an easy score-ordering problem, not an improvement with sample size. The remaining gap is known to arise from the policy-information restriction in this constructed example.

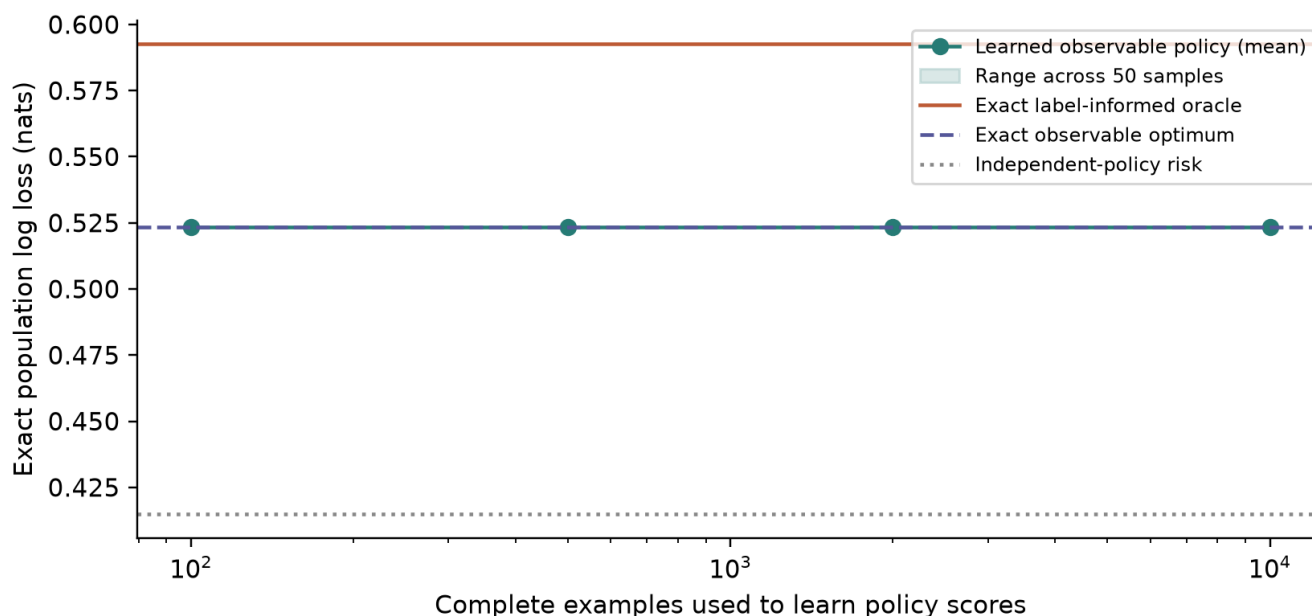


Figure 4. Known finite-distribution risks and learned observable-policy results. All 200 learned policies attain the observable optimum; this controlled construction separates information restriction from score-learning error.

A separate balanced binary example shows why a fixed mask histogram does not fix Bayes risk: assigning missingness by the label makes the mask reveal that label, giving zero policy-specific Bayes loss despite an unchanged half-budget. The model evaluated in that example remains fixed and need not exploit this new information. Such label-dependent policies are conceptual counterexamples, not recommendations for deployment.

The main experiment fitted 2,100 predictive models (1,620 linear and 480 tree fits). All linear optimizers reported success; the largest infinity-norm gradient was 1.01e-05, with one slightly above 0.00001. The independent output audit recomputed 13,320 validation/test prediction grids and 720 frozen selections; maximum metric discrepancy was 3.33e-16. Main code and protocol hashes matched their pre-run lock. Numerical checks passed for gradients, assignment enumeration, the integrated identity, smoothing bounds, and conditional allocation. These are internal consistency checks, not independent technical review. Elapsed fit timers include scheduling or interruption effects and are not used for computational-efficiency claims.

5.5. Study II: held-out model selection against shared policies

Table 7. Test maximum log loss over the same 61-condition policy library, averaged across ten configurations per dataset, at the prespecified $q=0.5$ and tolerance 0.03. Differences are nats; negative values favor shared-policy selection. Selector names describe the validation criterion, not different test targets.

Dataset	Nominal selector	Oracle selector	Shared selector	Shared - oracle	Shared - nominal
WDBC	0.1676	0.1634	0.1624	-0.0010	-0.0052
White Wine	0.5277	0.5237	0.5270	0.0033	-0.0006
Optical Digits	0.3264	0.3373	0.3282	-0.0092	0.0018
Page Blocks	0.1623	0.1646	0.1672	0.0025	0.0049
Vehicle	0.6649	0.6657	0.6608	-0.0049	-0.0041
Steel Faults	0.6787	0.6792	0.6795	0.0003	0.0008

The equally weighted primary difference is -0.0015 nats: shared-policy selection has lower worst-library loss on average, with lower means on 3 of six datasets. Across 60 configurations it wins 27, ties 21, and loses 12 using a 0.000001-nat comparison threshold. The median dataset difference is -0.0004; leave-one-dataset-out means range from -0.00245 to +0.00004 nats. Omitting Optical Digits changes the mean difference to +0.00004 nats, so the average advantage is sensitive to dataset composition. These numbers describe effect size and heterogeneity, not a significance test or an application-specific importance threshold.

Ordinary nominal-loss selection is a necessary control. Against it, shared-policy selection has 0.0004 nats lower average worst-library loss, with 22 configuration wins, 30 ties, and 8 losses. A comparison with oracle selection alone cannot establish added value beyond that simpler choice.

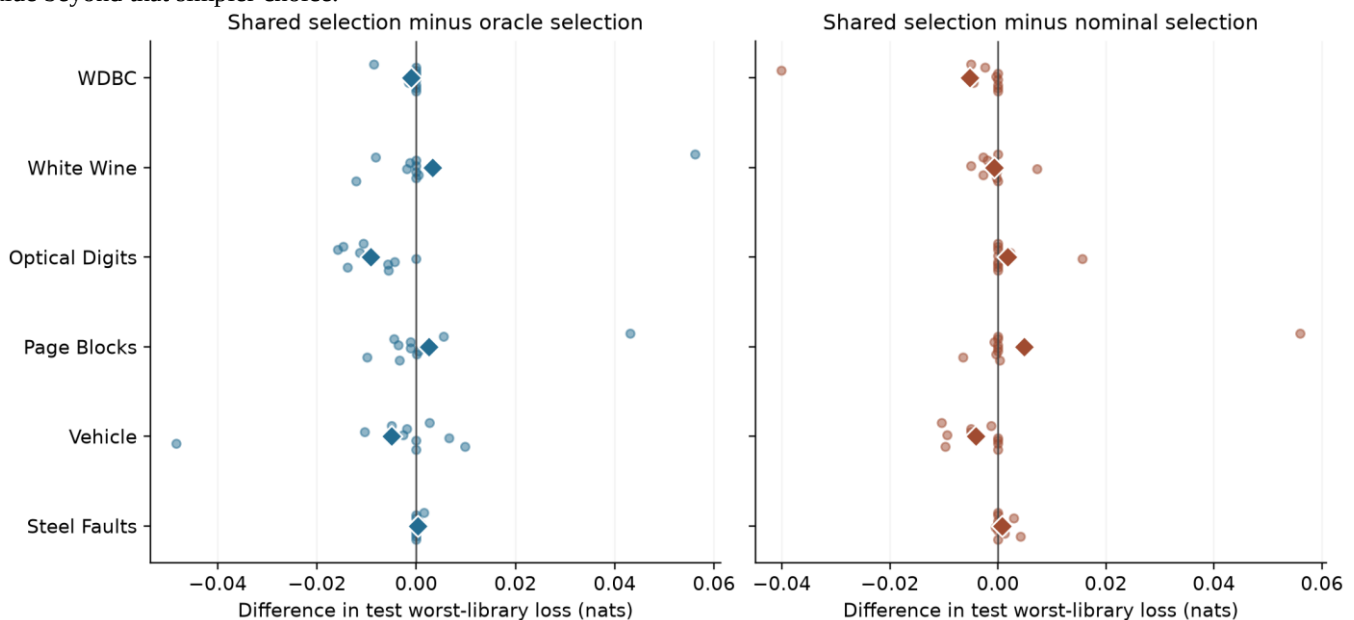


Figure 5. Study II primary paired differences in held-out worst-library log loss. Small points are overlapping configurations; diamonds are dataset means. Both comparisons use the same candidate eligibility rule and test library.

The nominal constraint fails on test in 13/60 shared-policy selections, 13/60 oracle selections, and 10/60 nominal selections at the primary tolerance. The validation condition is therefore a selection rule rather than a guaranteed test constraint. The appendix reports nominal losses, retrospective regrets, estimator-family transfers, and all twelve rate/tolerance settings.

With pseudorandom tie-breaking, the dataset-averaged differences are -0.0017 nats against oracle selection and -0.0006 against nominal selection. Ten permutations per configuration measure this particular tie sensitivity and do not create additional independent datasets.

Study II fitted 480 predictive models and 1,680 loss-increment regressors. All linear optimizers reported success. Independent recalculation paths checked 307,440 shared-policy risk entries, 21,600 exact-count masks, and 3,600 saved selections; the maximum numerical discrepancy was $2e-15$. Both study protocol locks match their executed fitting code. These are internal verification results.

6. DISCUSSION

The exact binary example establishes a decision consequence of policy information: oracle-based model selection can reject a useful input because its uncertainty set permits outcome-informed assignments. That is different from discovering that two risk envelopes have different values. The example does not invalidate the oracle target; it shows that a conservative target and an operational target can favor different predictors.

Study II tests this consequence using fresh datasets and a shared class of observable score rules. Its primary dataset-average difference of -0.0015 nats and nominal-selector comparison of -0.0004 nats delimit the empirical claim. Only three dataset means favor shared selection over the oracle selector, and the average advantage reverses without Optical Digits. The improvement over ordinary nominal selection is small. Thus the new experiment does not establish a reliably superior selector; a simpler nominal

criterion remains a serious alternative. Shared-policy selection directly targets the evaluated policy library, so better performance on that target would not establish robustness to mechanisms outside it.

The contribution is an empirical model-selection assessment connected to an exact explanatory example and a reproducible paired-loss audit. The underlying sorting, expected-shortfall, Gini, conditional-regression, constrained-allocation, and transportation tools are established. The new evidence concerns their joint use for this specified selection problem. Study I's negative robust-training comparison remains relevant: a new optimizer is not required for the evaluation question, and the tested robust comparator did not establish superiority over simple blends.

The primary practical recommendation is to state the policy information and selection target explicitly, and include ordinary nominal selection as a control. An oracle envelope is a conservative reference. A learned observable library is an achieved stress assessment. Neither should be silently described as the other. A common library makes the uncertainty class comparable across models, but the worst library member can still differ for each model.

Limitations

All missingness is imposed on complete public records. Feature blocks are documented groups or controlled random groups, not independently validated sensors or measured collection costs. Some variables are transformations of other variables, so removing a group need not remove its underlying information. No real collection-policy change, causal outcome intervention, or identification from naturally incomplete development data is demonstrated.

The twelve datasets are convenience benchmarks, with only six in the new selection test. WDBC and Vehicle have small selection/test partitions. Page Blocks contains blocks from 54 documents without released document identifiers, so document-disjoint splitting cannot be verified. Vehicle views may be related, and the original Satellite data can have spatial dependence. Exact duplicate grouping does not solve those provenance dependencies or near-duplicates. Repeated official test sets and overlapping splits preclude interpreting configurations as independent studies.

The fixed candidate family has moderate-size linear and boosting models and simple probability blends. Training remains at half missingness. Study I's neural experiment is a finite-budget objective adaptation, not a reproduction of the original MIRRAMS architecture. No comparison here establishes superiority over full StableMiss, NeuMISE, tabular foundation models, or a fully optimized acquisition system.

Policy estimation error and finite-library coverage remain unresolved. The score library is trained on a separate sample, but it is still constructed from the candidate predictors and evaluated as a specified robustness target. Alternative estimator families and tie permutations test limited forms of sensitivity; they do not establish coverage of real mechanisms. The independent-assignment condition is evaluated in expectation, whereas score policies use exact-count batches. Source row order can influence primary tie-breaking, and randomized tie sensitivity is reported separately.

Log loss is unbounded in general, and no uniform finite-sample generalization certificate is claimed. Empirical hindsight regret and validation nominal constraints are not deployment guarantees. The exact population example proves possibility, not prevalence. The targeted literature comparison supports precise attribution and a scoped contribution; it cannot certify exhaustive priority or predict editorial significance.

7. CONCLUSION

An oracle missingness envelope can be mathematically valid while selecting a poorer model for an information-restricted collection process. An exact example establishes that possibility. A separately specified six-dataset extension then compares oracle, nominal, and shared-observable-policy selection with common candidates, policy assignments, and held-out evaluation. At its primary setting, shared-policy selection differs from oracle selection by -0.0015 nats and from nominal selection by -0.0004 nats in dataset-averaged worst-library loss. The average advantage over oracle selection is sensitive to dataset composition, and the comparison with nominal selection is small. These findings support reporting the selection target and comparator explicitly; they do not establish a generally better selector. The contribution is the controlled assessment and its reproducible evidence.

DATA AND CODE AVAILABILITY

The accompanying reproducibility archive contains the public dataset archives and their source hashes, frozen protocol, deviations, executable experiments, verification scripts, per-configuration metrics, selected predictions, figure-generation code, and manuscript source. Original datasets are attributed in references 17-22 and 29-34 and retain their CC BY 4.0 licenses; their use here adds imposed masks and derived experimental results. Computations used local CPU resources and no paid services. The archive includes the full Study II candidate predictions and policy scores, both protocol locks, all selections, secondary diagnostics, and manuscript-generation code. The larger original Study I prediction grid can be regenerated from the supplied commands; selected Study I predictions are included directly.

REFERENCES

- [1] Rockenschaub, P., Xian, Z., Zamanian, A., Piperno, M., Ciora, O.-A., Pachl, E., and Ahmidi, N. (2024). Robust prediction under missingness shifts. arXiv:2406.16484.
- [2] Zhu, Y., Yuan, J., Jiang, B., Lin, T., Jin, H., Wang, X., and Zhou, C. (2023). Prediction with Incomplete Data under Agnostic Mask Distribution Shift. arXiv:2305.11197.
- [3] Lee, J., Kang, M., and Kim, D. (2025). MIRRAMS: Learning Robust Tabular Models under Unseen Missingness Shifts. arXiv:2507.08280, version 2.
- [4] Nguyen, M., Karaman, B. K., Kim, H., Wang, A. Q., Liu, F., and Sabuncu, M. R. (2025). Knockout: A simple way to handle missing inputs. Transactions on Machine Learning Research.
- [5] Globerson, A., and Roweis, S. (2006). Nightmare at Test Time: Robust Learning by Feature Deletion. Proceedings of ICML. DOI:10.1145/1143844.1143889.
- [6] Khan, F. A., Herasymuk, D., Protsiv, N., and Stoyanovich, J. (2025). Still More Shades of Null: An Evaluation Suite for Responsible Missing Value Imputation. PVLDB 18(9), 2899-2913. DOI:10.14778/3746405.3746416.
- [7] Rockafellar, R. T., and Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. Journal of Banking and Finance 26(7), 1443-1471. DOI:10.1016/S0378-4266(02)00271-6.
- [8] Sordo, M. A., de Souza, M. C., and Suarez-Llorens, A. (2016). Testing variability orderings by using Gini's mean differences. Statistical Methodology 32, 63-76. DOI:10.1016/j.stamet.2016.03.001.
- [9] Guastaroba, G., Mansini, R., Ogryczak, W., and Speranza, M. G. (2020). Enhanced index tracking with CVaR-based ratio measures. Annals of Operations Research 292, 883-931. DOI:10.1007/s10479-020-03518-7.
- [10] Athey, S., and Wager, S. (2021). Policy Learning With Observational Data. Econometrica 89(1), 133-161. DOI:10.3982/ECTA15732.
- [11] Kitagawa, T., and Tetenov, A. (2018). Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice. Econometrica 86(2), 591-616. DOI:10.3982/ECTA13288.
- [12] Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1993). Network Flows: Theory, Algorithms, and Applications. Prentice Hall.
- [13] Boyd, S., and Vandenberghe, L. (2004). Convex Optimization. Cambridge University Press.
- [14] Zhou, H., Balakrishnan, S., and Lipton, Z. C. (2023). Domain Adaptation under Missingness Shift. arXiv:2211.02093, version 3.
- [15] Choi, Y. S., Jeanselme, V., Elias, P., and Joshi, S. (2026). ICYM2I: The illusion of multimodal informativeness under missingness. arXiv:2505.16953, version 3.
- [16] Jang, S., Eun, J., Yi, H., and Choi, C. (2026). When Does Train-Time Masking Help? A Leakage-Controlled Robustness Audit of Tabular Classifiers Under Feature Missingness. IEEE Access 14, 84378-84398. DOI:10.1109/ACCESS.2026.3695538.
- [17] Lohweg, V. (2012). Banknote Authentication Dataset. UCI Machine Learning Repository. DOI:10.24432/C55P57. CC BY 4.0.
- [18] Hopkins, M., Reeber, E., Forman, G., and Suermondt, J. (1999). Spambase Dataset. UCI Machine Learning Repository. DOI:10.24432/C53G6X. CC BY 4.0.
- [19] Bock, R. (2004). MAGIC Gamma Telescope Dataset. UCI Machine Learning Repository. DOI:10.24432/C52C8B. CC BY 4.0.
- [20] Alpaydin, E., and Alimoglu, F. (1996). Pen-Based Recognition of Handwritten Digits Dataset. UCI Machine Learning Repository. DOI:10.24432/C5MG6K. CC BY 4.0.
- [21] Dry Bean Dataset. (2020). UCI Machine Learning Repository. DOI:10.24432/C50S4B. CC BY 4.0. Associated data-collection paper: Koklu, M., and Ozkan, I. A. (2020), Multiclass classification of dry beans using computer vision and machine learning techniques.
- [22] Srinivasan, A. (1993). Statlog (Landsat Satellite) Dataset. UCI Machine Learning Repository. DOI:10.24432/C55887. CC BY 4.0.
- [23] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825-2830.
- [24] Alzhrani, K. M. (2026). Decision Reliability Profiles for Auditing Targeted Case-Removal Sensitivity in AI Model Comparisons Under Additive Metrics. Algorithms 19(8), 676. DOI:10.3390/a19080676.
- [25] Gadgil, S., Covert, I., and Lee, S.-I. (2024). Estimating Conditional Mutual Information for Dynamic Feature Selection. International Conference on Learning Representations. arXiv:2306.03301v3.
- [26] Valancius, M., Lennon, M., and Oliva, J. (2024). Acquisition Conditioned Oracle for Nongreedy Active Feature Acquisition. Proceedings of Machine Learning Research 235, 48957-48975.
- [27] Charusaie, M.-A., and Samadi, S. (2024). A Unifying Post-Processing Framework for Multi-Objective Learn-to-Defer Problems. Advances in Neural Information Processing Systems 37.
- [28] Gao, R., and Kleywegt, A. J. Data-Driven Robust Optimization with Known Marginal Distributions. Author-hosted working paper, accessed September 8, 2026.
- [29] Wolberg, W., Mangasarian, O., Street, N., and Street, W. (1993). Breast Cancer Wisconsin (Diagnostic) Dataset. UCI Machine Learning Repository. DOI:10.24432/C5DW2B. CC BY 4.0.
- [30] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Wine Quality Dataset. UCI Machine Learning Repository. DOI:10.24432/C56S3T. CC BY 4.0. White-wine subset used here.
- [31] Alpaydin, E., and Kaynak, C. (1998). Optical Recognition of Handwritten Digits Dataset. UCI Machine Learning Repository. DOI:10.24432/C50P49. CC BY 4.0.
- [32] Malerba, D. (1994). Page Blocks Classification Dataset. UCI Machine Learning Repository. DOI:10.24432/C5J590. CC BY 4.0.
- [33] Mowforth, P., and Shepherd, B. Statlog (Vehicle Silhouettes) Dataset. UCI Machine Learning Repository. DOI:10.24432/C5HG6N. CC BY 4.0.
- [34] Buscema, M., Terzi, S., and Tastle, W. (2010). Steel Plates Faults Dataset. UCI Machine Learning Repository. DOI:10.24432/C5J88N. CC BY 4.0.

APPENDIX A. ADDITIONAL NUMERICAL RESULTS

A.1. Paired differences by dataset and tolerance

Table A1. Robust minus blended linear losses. Standard deviations describe ten overlapping split/panel results; they are not standard errors or confidence intervals.

Dataset	Tolerance	Nominal mean	Nominal SD	Oracle mean	Oracle SD
Banknote	0.0000	-0.0006	0.0010	-0.0004	0.0024
Banknote	0.0100	-0.0030	0.0041	0.0001	0.0021
Banknote	0.0300	-0.0093	0.0066	0.0017	0.0034
Banknote	0.1000	-0.0256	0.0220	0.0021	0.0018
Dry Bean	0.0000	-0.0002	0.0003	-0.0006	0.0017
Dry Bean	0.0100	-0.0004	0.0002	0.0009	0.0002
Dry Bean	0.0300	-0.0004	0.0002	0.0009	0.0002
Dry Bean	0.1000	-0.0004	0.0002	0.0009	0.0002
MAGIC	0.0000	0.00001	0.0003	0.0065	0.0046
MAGIC	0.0100	-0.0003	0.00004	0.0012	0.0001
MAGIC	0.0300	-0.0003	0.00004	0.0012	0.0001
MAGIC	0.1000	-0.0003	0.00004	0.0012	0.0001
Pendigits	0.0000	0.0008	0.0026	0.0024	0.0075
Pendigits	0.0100	-0.0035	0.0014	0.0079	0.0086
Pendigits	0.0300	-0.0067	0.0044	0.0017	0.0094
Pendigits	0.1000	-0.0217	0.0111	0.0055	0.0098
Satellite	0.0000	0.0014	0.0042	0.0098	0.0155
Satellite	0.0100	-0.0020	0.0005	0.0035	0.0010
Satellite	0.0300	-0.0021	0.0002	0.0036	0.0006
Satellite	0.1000	-0.0021	0.0002	0.0036	0.0006
Spambase	0.0000	0.0010	0.0053	0.0130	0.0193
Spambase	0.0100	-0.0032	0.0029	0.0039	0.0079
Spambase	0.0300	-0.0063	0.0088	0.0027	0.0052
Spambase	0.1000	-0.0130	0.0067	0.0044	0.0024

A.2. Four-pattern assignment

Table A2. Four-pattern test log losses in nats with fixed quotas; one seed/panel per dataset.

Dataset	Model	Independent	Minimum	Learned	Maximum
Banknote	linear	0.2358	0.0971	0.3403	0.4236
Banknote	tree	0.1057	0.0172	0.1987	0.2434
Spambase	linear	0.2862	0.1373	0.2850	0.4669
Spambase	tree	0.1854	0.0943	0.2259	0.3054
MAGIC	linear	0.4558	0.4141	0.4638	0.4977
MAGIC	tree	0.3212	0.2550	0.3294	0.3905
Pendigits	linear	0.5014	0.3272	0.5903	0.7005
Pendigits	tree	0.2280	0.1797	0.2469	0.2845
Dry Bean	linear	0.2327	0.2088	0.2358	0.2577
Dry Bean	tree	0.2293	0.2170	0.2288	0.2418
Satellite	linear	0.3887	0.3240	0.3925	0.4585
Satellite	tree	0.3131	0.2886	0.3150	0.3386

A.3. Neural checkpoint and nominal loss

Table A3. Supplementary neural checkpoint selection and stopping. Nominal denotes independent-assignment log loss in nats. Block and complete denote the two MIRRAMS-objective starting-input adaptations.

Dataset	Objective	Nominal	Best epoch	Last epoch
Banknote	independent	0.1188	965	1000
Banknote	block	0.1365	915	1000
Banknote	complete	0.1486	985	1000
Dry Bean	independent	0.1925	1000	1000
Dry Bean	block	0.2025	980	1000
Dry Bean	complete	0.2280	970	1000
MAGIC	independent	0.3323	360	460
MAGIC	block	0.3300	990	1000
MAGIC	complete	0.3660	385	485
Pendigits	independent	0.1537	490	590
Pendigits	block	0.1543	1000	1000
Pendigits	complete	0.2046	800	900
Satellite	independent	0.2871	760	860
Satellite	block	0.2942	1000	1000
Satellite	complete	0.3330	930	1000
Spambase	independent	0.2115	165	265
Spambase	block	0.2119	340	440
Spambase	complete	0.2203	245	345

A.4. Source-mask diagnostics

Table A4. Mean test log losses in nats. Outcome-associated source masking creates a deliberately injected shortcut. Extra deletion reduces several losses but does not eliminate sensitivity. Aligned and reversed policies preserve counts; they are labeled stress conditions.

Dataset	Training	Independent test	Aligned test	Reversed test
Banknote	source	0.2677	0.1184	0.4001
Banknote	extra	0.2249	0.1611	0.2790
Banknote	independent	0.1901	0.2091	0.1611
Spambase	source	0.5447	0.2775	0.8074
Spambase	extra	0.4960	0.3008	0.6868
Spambase	independent	0.3305	0.3434	0.3162
MAGIC	source	0.7327	0.2916	1.1713
MAGIC	extra	0.6658	0.3207	1.0094
MAGIC	independent	0.4596	0.4639	0.4558
Pendigits	source	0.6278	0.4706	0.7897
Pendigits	extra	0.5756	0.5028	0.6511
Pendigits	independent	0.5244	0.5418	0.5060
Dry Bean	source	0.3025	0.1772	0.4250
Dry Bean	extra	0.2751	0.1874	0.3613
Dry Bean	independent	0.2266	0.2254	0.2281
Satellite	source	0.5436	0.3148	0.7762
Satellite	extra	0.4839	0.3232	0.6456
Satellite	independent	0.3896	0.3866	0.3931

A.5. Reproducibility and scope

The archive README gives exact commands, package versions, dataset attribution, retained prediction keys, and verification steps. Full-grid CSV results are descriptive records and must not be used to replace frozen selections. Reproducing the largest independent prediction audit requires regenerating full-grid binary predictions; selected prediction files are included for direct inspection. No study outcome has been externally reviewed as part of these computations.

A.6. Average removal loss and integrated sensitivity

Table A5. Dataset means across ten configurations. The mean increment measures average removal loss; the integrated premium measures dispersion of those increments, in nats integrated over the unit interval of budgets. These quantities describe the fitted predictor and need not rank datasets or backbones identically.

Dataset	Model	Full-input loss	Missing-input loss	Mean increment	Integrated premium
Banknote	linear	0.0620	0.3185	0.2565	0.1267
Banknote	tree	0.0285	0.2555	0.2270	0.0997
Spambase	linear	0.2966	0.3644	0.0678	0.0905
Spambase	tree	0.1385	0.2059	0.0674	0.0540
MAGIC	linear	0.4556	0.4635	0.0079	0.0325
MAGIC	tree	0.2966	0.3554	0.0587	0.0606
Pendigits	linear	0.4225	0.6264	0.2038	0.1640
Pendigits	tree	0.1881	0.2733	0.0852	0.0498
Dry Bean	linear	0.2236	0.2295	0.0058	0.0138
Dry Bean	tree	0.2373	0.2388	0.0015	0.0103
Satellite	linear	0.3787	0.4005	0.0218	0.0472
Satellite	tree	0.3010	0.3062	0.0052	0.0179

A.7. Validation-to-test tolerance transfer

Table A6. Counts of test nominal-loss tolerance violations, using a 0.000001-nat numerical margin, for every selection family. Each family uses its own independent backbone reference.

Tolerance	Robust linear	Blended linear	Blended trees
0.0000	10/60	19/60	15/60
0.0100	5/60	8/60	7/60
0.0300	1/60	7/60	1/60
0.1000	1/60	5/60	0/60

A.8. Partition sizes and implementation details

Table A7. Row-count ranges over five seeds; feature panels reuse the same partitions. Excluded rows are exact feature-vector overlaps between official development and test data.

Dataset	Training rows	Validation rows	Test rows	Excluded development rows
Banknote	818-825	272-279	270-276	0
Spambase	2718-2820	895-981	884-988	0
MAGIC	11408-11423	3798-3809	3798-3807	0
Pendigits	5995	1499	3498	0
Dry Bean	8159-8173	2719-2726	2719-2729	0
Satellite	3548	887	2000	0

Split seeds are 314+s for the first grouped split and, for datasets without official tests, 7919+s for the development split, with s=0,...,4. Optional panels use NumPy default random generators seeded by 1701+p for panel p=0 or 1. Panels can overlap; they are not disjoint feature blocks across the two panel settings. Constant or nearly constant training standard deviations below 10⁻¹² are set to one. Ties in validation selection are resolved by worst loss, then nominal loss, then the deterministic candidate identifier. Learned-policy score ties use a separate seed 9182+s. The retained split manifests record original row indices, class order, standardization parameters, and optional-feature indices. All dependency versions are included in the archive.

A.9. Study II rate, tolerance, and estimator sensitivity

Table A8. Dataset-averaged shared-selector minus comparator differences in test worst-library loss. All selectors face the same full test library; HGB and ridge selectors use their smaller libraries for selection. These settings were specified before Study II fits; the primary row is rate 0.5, tolerance 0.03.

Rate	Tolerance	Vs oracle	Vs nominal	Vs HGB	Vs ridge
0.2500	0.0000	0.0000	0.0000	0.0000	0.0000
0.2500	0.0100	-0.0037	-0.0011	-0.0002	-0.0009
0.2500	0.0300	-0.0065	-0.0003	-0.0002	-0.0001
0.2500	0.1000	-0.0119	-0.0003	-0.0002	-0.0001
0.5000	0.0000	0.0000	0.0000	0.0000	0.0000
0.5000	0.0100	0.0003	-0.0014	-0.00002	-0.0001
0.5000	0.0300	-0.0015	-0.0004	-0.00002	0.00002
0.5000	0.1000	-0.0040	-0.0005	-0.00002	-0.00003
0.7500	0.0000	0.0000	0.0000	0.0000	0.0000
0.7500	0.0100	-0.0003	-0.0021	-0.0003	-0.0004
0.7500	0.0300	-0.0020	-0.0020	-0.0003	-0.0004
0.7500	0.1000	-0.0020	-0.0020	-0.0003	-0.0004

Table A9. Dataset-averaged primary-setting outcomes. Regret uses the best test worst-library candidate within the validation-eligible set and is not a usable selection oracle.

Selector	Nominal loss	Worst-library loss	Hindsight regret
nominal	0.4022	0.4212	0.0087
oracle	0.4130	0.4223	0.0098
shared	0.4062	0.4208	0.0083
hgb	0.4056	0.4209	0.0083
ridge	0.4055	0.4208	0.0083

Table A10. Transfer across the two learned score families at the primary setting. Feature and uniform policies are shared between libraries. The difference is test loss under the indicated library; it measures limited estimator-family sensitivity, not transfer to real collection mechanisms.

Dataset	Selection library	Test library	Minus oracle selector
WDBC	hgb	ridge	-0.0010
White Wine	hgb	ridge	0.0024
Optical Digits	hgb	ridge	-0.0126
Page Blocks	hgb	ridge	0.0024
Vehicle	hgb	ridge	-0.0047
Steel Faults	hgb	ridge	0.0003
WDBC	ridge	HGB	-0.0010
White Wine	ridge	HGB	0.0034
Optical Digits	ridge	HGB	-0.0102
Page Blocks	ridge	HGB	0.0026
Vehicle	ridge	HGB	-0.0058
Steel Faults	ridge	HGB	0.0003

A.10. Exact model-selection example grid

Table A11. Prespecified binary example grid; A is the full conditional predictor when S is available and B always predicts one-half. All entries integrate the exact finite distribution.

P(S=Y)	Missing rate	Oracle choice	Observable choice	Observable regret
0.6000	0.2500	B	A	0.0151
0.6000	0.5000	B	A	0.0101
0.6000	0.7500	B	A	0.0050
0.7000	0.2500	B	A	0.0617
0.7000	0.5000	B	A	0.0411
0.7000	0.7500	B	A	0.0206
0.8000	0.2500	A	A	0.0000
0.8000	0.5000	B	A	0.0964
0.8000	0.7500	B	A	0.0482
0.9000	0.2500	A	A	0.0000
0.9000	0.5000	A	A	0.0000
0.9000	0.7500	B	A	0.0920
0.9500	0.2500	A	A	0.0000
0.9500	0.5000	A	A	0.0000
0.9500	0.7500	A	A	0.0000

Table A12. Two observed groups of equal probability have optional-bit accuracies 0.6 and 0.9. Both have balanced outcomes. Candidate probabilities shrink the correct group-specific posterior toward one-half with weights 0, 0.25, 0.5, 0.75, and 1. The observable policy sees the group. This is a finite-family illustration, not a new fitted-method comparison.

Missing rate	Oracle weight	Observable weight	Observable regret
0.2500	0.7500	1.0000	0.0098
0.5000	0.0000	1.0000	0.0101
0.7500	0.0000	1.0000	0.0050

Copyright & License:

© Authors retain the copyright of this article. This work is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.