WHAT DOES A PREDICTIVE GAIN CERTIFY? AUDITING INFORMATION CLAIMS UNDER APPROXIMATE INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

A better predictor need not use new information: it may simply make better use of information already available. We study when a held-out predictive gain supports a claim of additional conditional information. Our audit connects three requirements: attributing the gain, validating the baseline's approximation error, and accounting for uncertainty in the final decision. We prove that probability reports and labels alone cannot distinguish some informative and non-informative problems. For a known categorical baseline, we derive an exact reverse-to-forward KL conversion requiring only its smallest probability and one scalar root, avoiding exponential subset enumeration. A failure-aware audit incorporates uncertain error bounds and supports both betting and fixed-sample confidence statements. For sampled categorical reports, simultaneous intervals and a simplex constraint avoid the universal-bound failure caused by unseen classes. On natural WDBC, Digits, and Wine labels, an independent witness sample yields positive simultaneous information bounds in six of nine fixed analyses; all three Wine analyses remain inconclusive. A learned-representation audit checks attribution on continuous 64-dimensional inputs. A deterministic deeper layer passes the predictive-gain test in every 8,000-case repetition despite having zero additional information; the corrected audit makes no positive information claim. With an added informative signal, the betting audit detects it in 98% of repetitions, compared with 100% for conditional randomization using its exact feature sampler.

1 INTRODUCTION

A representation can improve prediction without adding information. A linear model given H may benefit from $G = H^2$, although this feature is fully determined by H. It makes an existing relationship easier to learn. This distinction between available and usable information is established (Xu et al., 2020). The question is what the improvement establishes about the input.

Approximate Bayesian inference creates the same ambiguity: recovering a missed posterior region can improve prediction without adding an observation. Test data confirm the gain but cannot attribute it to new evidence (Huggins et al., 2020).

We ask: *what additional guarantee turns a predictive comparison into evidence that an added input contains information?* The established relation between log score, information, and approximation error gives an asymmetry: a lower information bound needs an upper bound on the baseline's error (Gneiting and Raftery, 2007; Barber and Agakov, 2003). Inaccuracy in the improved predictor makes the bound conservative; inaccuracy in the baseline can inflate apparent information.

Three requirements define the audit. **Attribution:** separate additional information from improved approximation. **Directional validation:** convert a supplied reverse-KL budget into forward predictive error, including Monte Carlo estimation. **Failure-aware accounting:** include baseline error, failures of its bound, and audit uncertainty in the final confidence statement. A separation theorem shows why accurate sampling diagnostics cannot substitute for validation. Exact binary and categorical constructions provide the conversion; a finite-sample theorem provides the audit. Experiments separate approximation improvement, additional information, and uncertainty costs.

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.

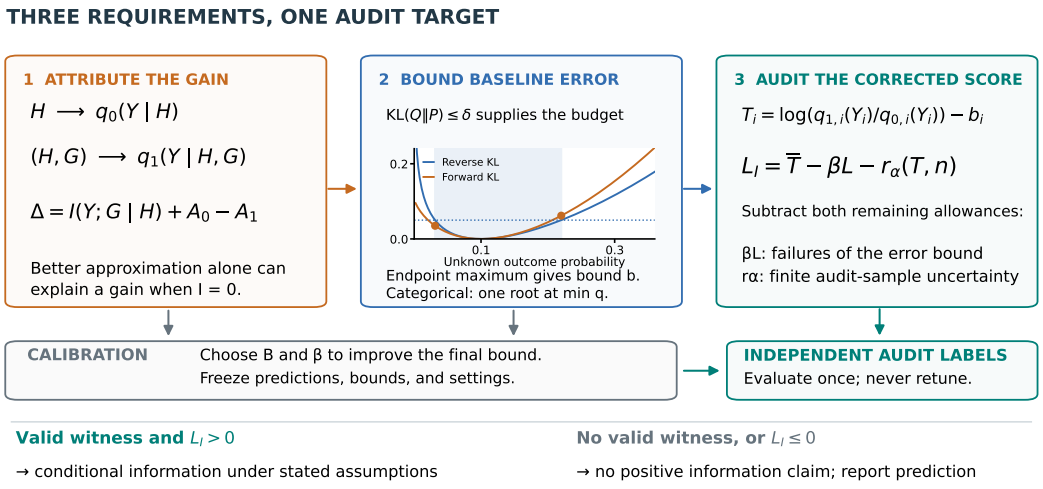

Figure 1: **From predictive gain to an information claim.** Y is the outcome, H the existing input, and G the added input. Frozen predictors $q_0(Y \mid H)$ and $q_1(Y \mid H, G)$ report baseline and candidate probabilities. Δ is expected log-score gain; $I = I(Y; G \mid H)$ is conditional information; A_0, A_1 are the predictors' expected forward-KL errors relative to their true conditional laws. P is the target posterior, Q its approximation, and δ their supplied reverse-KL budget; KL denotes Kullback–Leibler divergence. The binary inset uses approximate probability $u = .1$ and $\delta = .05$; q is a known categorical baseline vector and $\min q$ its smallest entry. The bound b controls baseline predictive error. For audit case i, Y_i is its outcome, $q_{0,i}, q_{1,i}$ its reported distributions, b_i its error bound, and T_i its corrected log-score gain. $\overline{T}$ averages n independent cases. β bounds the error-bound failure probability; $L = \log(1/\epsilon)$ bounds baseline log loss, with probability floor $\epsilon > 0$. The sampling allowance $r_\alpha(T, n)$ has error probability at most α; L_I is the resulting untruncated information lower bound. Calibration selects log-score cap B and β before the independent audit.

2 Prediction Gain and Information Are Different Targets

Let Y be an outcome with finitely many possible values, H the input already available, and G an additional input. In a reliability application, Y can be the indicator that a frozen classifier is correct; H contains its output and other available diagnostics, and G contains a proposed uncertainty signal. In a Bayesian application, H can contain the observations already acquired and G a proposed additional measurement.

Write $p_0(y \mid H) = P(Y = y \mid H)$ and $p_1(y \mid H, G) = P(Y = y \mid H, G)$. Let $q_0(y \mid H)$ and $q_1(y \mid H, G)$ be predictors fixed before evaluation, each assigning positive probability to every outcome. All expectations below refer to the stated evaluation population and are assumed finite. The log score is the logarithm of the probability assigned to the observed outcome; a higher score is better. We use natural logarithms, giving units called nats. Define

$$D = \log \frac{q_1(Y \mid H, G)}{q_0(Y \mid H)}, \qquad \Delta = \mathbb{E}[D], \tag{1}$$
$$A_0 = \mathbb{E}\,\mathrm{KL}(p_0(\cdot \mid H) \| q_0(\cdot \mid H)), \quad A_1 = \mathbb{E}\,\mathrm{KL}(p_1(\cdot \mid H, G) \| q_1(\cdot \mid H, G)).$$

Here Δ is the average gain over the evaluation population. Conditional mutual information, $I(Y; G \mid H)$, measures how much knowing G reduces uncertainty about Y once H is known. The terms A_0, A_1 are the predictors' expected excess log losses relative to the true conditional probabilities. They use forward KL, $\mathrm{KL}(p\|q) = \sum_y p_y \log(p_y/q_y)$. Reversing its arguments generally changes its value.

Proposition 1 (Established log-score decomposition)**.** *With $I = I(Y; G \mid H)$,*

$$\Delta = I + A_0 - A_1. \tag{2}$$

Consequently, if $A_0 \le U_0$, then $I \ge \Delta - U_0$. If also $A_1 \le U_1$, then $I \le \Delta + U_1$.

Table 1: **A stronger conclusion requires stronger evidence.** The same predictive gain supports different claims depending on the available error control.

Claim	Required evidence	What remains outside the claim
Predictive improvement	Valid lower bound $L_\Delta > 0$ for frozen predictors	Additional information; ranking or deployment utility
Conditional information	Valid baseline witness and $L_I > 0$	Model correctness if the witness is only model-relative
Deployment authorization	A separately declared gate with selection and multiplicity control	Arbitrary distribution shift or untested performance metrics

This follows by adding and subtracting $\log p_0(Y \mid H)$ and $\log p_1(Y \mid H, G)$; it is the conditional form of the proper log-score risk decomposition (Gneiting and Raftery, 2007). The decomposition is also related to calibration and refinement of proper scores (Bröcker, 2009). For an exact baseline, the lower bound is the familiar variational information bound (Barber and Agakov, 2003). We do not claim either identity as new.

The two approximation errors have different effects. A poor new predictor reduces Δ and therefore makes a lower information certificate harder to obtain. A poor baseline can increase Δ and therefore creates the opposite problem. Matching model capacity and tuning budgets may improve the fairness of a comparison, but does not provide a numerical bound on A_0.

A lower confidence bound L_Δ for Δ and a valid upper bound U_0 imply

$$L_I = \max\{0, L_\Delta - U_0\} \le I. \tag{3}$$

If the two bounds have failure probabilities α and γ, their joint failure probability is at most $\alpha + \gamma$; independence is unnecessary. If no useful U_0 is available, a positive L_Δ certifies a predictive improvement only. It does not test $Y \perp G \mid H$. This distinction is consistent with the difficulty of unrestricted conditional-independence testing (Shah and Peters, 2020) and information estimation (McAllester and Stratos, 2020).

Proposition 2 (Predictive reports do not identify information)**.** *There exist a conditional-independence null and an informative alternative with the same frozen predictor functions and the same distribution of their reported probabilities and labels. Every test using only these reports has equal rejection probability under the two distributions, for every sample size.*

For a fair binary G, take $q_0 = 1/2$, $q_1 = 1/2 + t(2G - 1)$, and $Y \mid G \sim \mathrm{Bernoulli}(q_1)$. If H is independent, $I = \mathrm{kl}(1/2 + t \| 1/2) > 0$; if $H = G$, $I = 0$. The score reports are unchanged. This is a limitation of score-only evidence, not of conditional-independence tests using structural assumptions on the full inputs.

3 Why small diagnostic errors are not enough

Forward and reverse KL penalize different errors. In particular, assigning very little probability to a rare event can be inexpensive in reverse KL and expensive in expected log loss (Sason and Verdú, 2016; Huggins et al., 2020). The following example combines that asymmetry with increasing evaluation precision.

Theorem 1 (Joint diagnostic separation)**.** *For every integer $n > 1$, let H, G be constant, $Y \sim \mathrm{Bernoulli}(p_n)$, $p_n = n^{-1/4}$, $q_{0,n} = \exp(-\sqrt{n})$, and $q_{1,n} = p_n$. Then $I(Y; G \mid H) = 0$, while*

$$\mathrm{kl}(q_{0,n} \| p_n) \sim n^{-1/4}, \qquad \Delta_n \sim n^{1/4}, \qquad \sqrt{\mathrm{Var}(D_n)/n} \sim n^{-1/8}. \tag{4}$$

Here kl *denotes KL divergence between Bernoulli distributions. The last quantity is the actual standard deviation of the mean of n independent log-score differences.*

Both diagnostics therefore approach zero while predictive gain grows without limit, even though no information was added. The sampling standard error is not misestimated: it answers a different question. A test of $\Delta_n > 0$ is correctly detecting improvement in approximation. It becomes misleading only if interpreted as a test of $I > 0$.

Each value of n defines a different prediction problem. The example does not say that a fixed predictor becomes less accurate as more data are collected. It uses increasingly small baseline probabilities. A fixed probability floor prevents divergence, but still permits positive gains entirely due to approximation error. In Section 6, a fixed mixture model demonstrates that finite version.

4 BOUNDING THE BASELINE ERROR

We call a valid upper bound on the baseline error an *approximation witness*. It specifies how much of the observed improvement could be explained by a poor baseline. A witness must be a mathematical bound; a favorable diagnostic alone is not enough.

4.1 EXACT CONVERSION FOR BINARY PREDICTIONS

Suppose a validated calculation gives $\mathrm{kl}(u\|p) \le \delta$, where $u \in (0,1)$ is an approximate predictive probability and p is the true predictive probability. The required quantity is instead $\mathrm{kl}(p\|r)$, where $r \in (0,1)$ is the probability actually reported. Allowing $r \neq u$ accommodates clipping and Monte Carlo estimation.

For $\delta > 0$, let $p_-(u,\delta) < u < p_+(u,\delta)$ be the two solutions of $\mathrm{kl}(u\|p) = \delta$. Set both endpoints to u when $\delta = 0$.

Proposition 3 (Sharp binary witness)**.** *For any reported probability $r \in (0,1)$,*

$$b(r;u,\delta) = \max\{\mathrm{kl}(p_-(u,\delta)\|r), \mathrm{kl}(p_+(u,\delta)\|r)\} = \sup_{p:\,\mathrm{kl}(u\|p)\le\delta} \mathrm{kl}(p\|r). \tag{5}$$

The reverse-KL bound restricts the unknown probability p to an interval. Over that interval, forward KL is a convex function, so its largest value occurs at one of the two endpoints. This exact one-dimensional conversion specializes divergence-comparison ideas (Sason and Verdú, 2016) to the error needed in Equation (3).

For Bayesian inference, δ can come from a bound $\mathrm{KL}(Q(\theta \mid H)\|P(\theta \mid H)) \le \delta(H)$. Passing both distributions through the same predictive likelihood gives $\mathrm{kl}(u(H)\|p_0(H)) \le \delta(H)$ by data processing. This statement is about the posterior predictive of the specified model. It applies to the data-generating conditional distribution only if the model is correct, or if a separate discrepancy bound connects the two. The evidence lower bound (ELBO), commonly optimized in variational inference, does not by itself supply δ: its gap to the log model evidence is generally unknown. Validated inference bounds are a relevant source of additional control (Huggins et al., 2020).

4.2 AN EXACT CATEGORICAL BOUND WITH ONE SCALAR ROOT

Let q be a known, strictly positive K-class baseline with $K \ge 2$ and $m = \min_j q_j$. Write $v_+(m,\delta)$ for the upper solution of $\mathrm{kl}(m\|v) = \delta$ and $F_\delta(m) = \mathrm{kl}(v_+(m,\delta)\|m)$.

Theorem 2 (Least-mass categorical witness)**.** *For every $\delta \ge 0$,*

$$\sup_{p:\,\mathrm{KL}(q\|p)\le\delta} \mathrm{KL}(p\|q) = F_\delta(m). \tag{6}$$

An attaining distribution increases a least-probable class to $v_+(m,\delta)$ and rescales all remaining probabilities proportionally. Computing the bound needs an $O(K)$ minimum scan and one scalar root.

Proof sketch. At an extremum, Lagrange stationarity gives at most two distinct ratios p_j/q_j, reducing the problem to binary subset masses w. For the upper binary root v_+, let $d = v_+ - w$ and $t = \log\{v_+(1-w)/(w(1-v_+))\}$. Differentiation gives

$$F'_\delta(w) = -\frac{d}{w(1-w)} + \frac{t^2 v_+(1-v_+)}{d} \le 0,$$

because $d/\sqrt{v_+(1-v_+)w(1-w)} = 2\sinh(t/2) \ge t$. Complement symmetry handles the lower root. Every subset mass lies in $[m, 1-m]$, and the singleton of mass m attains the maximum. □

The proof specializes established divergence geometry (Harremoës and Vajda, 2011; Sason and Verdú, 2016) but removes the exponential partition search.

Theorem 3 (Estimated categorical witness)**.** *Let* $\mathrm{KL}(u\|p) \le \delta$ *and let simultaneous intervals* $[l_j, U_j]$ *cover the unknown sampling probabilities* u_j *with probability at least* $1-\beta$*. For a positive report* r*, set* $a_j = p_-(l_j, \delta)$ *and* $c_j = p_+(U_j, \delta)$*. Let* C *maximize the sum of chords of* $x\log(x/r_j)$ *on* $[a_j, c_j]$*, subject to* $\sum_j x_j = 1$*. Then*

$$b = \max\{0, \min\{\log(1/\min_j r_j), C, \log\max_j(c_j/r_j)\}\}$$

upper-bounds $\mathrm{KL}(p\|r)$ *with probability at least* $1-\beta$*. An infeasible interval box uses the universal bound instead.*

Binary data processing places p inside the interval box. Convexity bounds each KL term by its chord; sorting chord slopes solves the constrained maximum in $O(K\log K)$ time after endpoint conversion. With Clopper–Pearson intervals, b is strictly below the universal bound whenever at least two classes are observed, even if other classes are unseen. The supplement proves this property, gives an additional valid envelope whose inclusion recovers the exact known-vector limit, and states all endpoint conventions. Nonvacuity does not imply a small error bound.

4.3 When predictions are estimated from samples

Suppose the predictive probability is an average, $u(H) = \mathbb{E}_Q[\ell(\theta, H)]$, where each sampled likelihood value satisfies $0 \le \ell \le 1$. From S independent draws from Q, obtain $\widehat{u}$ and the interval

$$[u_L, u_U] = [\widehat{u} - t, \widehat{u} + t] \cap [0,1], \qquad t = \sqrt{\frac{\log(2/\beta)}{2S}}. \tag{7}$$

Conditional on H, this contains $u(H)$ with probability at least $1-\beta$. The endpoints $p_-(u,\delta)$ and $p_+(u,\delta)$ are increasing in u. Hence every p_0 compatible with this interval and the reverse-KL budget lies in

$$[a, c] = [p_-(u_L, \delta), p_+(u_U, \delta)]. \tag{8}$$

For $r = \mathrm{clip}(\widehat{u}, \epsilon, 1-\epsilon)$, set $b = \max\{\mathrm{kl}(a\|r), \mathrm{kl}(c\|r)\}$. With probability at least $1-\beta$, this bounds the forward KL of the *reported* prediction. Endpoint formulas also cover $u_L = 0$ and $u_U = 1$; details are in the supplement.

Equation (7) is not a valid substitute for a mixing bound when draws come from an arbitrary Markov chain. A chain-based procedure must provide its own valid interval. Increasing the number of dependent samples or observing a favorable convergence diagnostic does not establish the stated coverage (Vehtari et al., 2021).

5 Accounting for error bounds that sometimes fail

The previous construction can fail on some cases. Per-case conditional coverage implies the weaker unconditional failure assumption used below. Requiring every Monte Carlo interval to succeed would make its confidence level depend on the number of audit cases. Instead, we can bound how much these failures affect the average error.

Let U_0 denote fresh randomness used by the baseline and witness, with $U_0 \perp (Y, G) \mid H$. The predictor with the additional input may use H, G, U_0 and additional randomness independent of Y given those quantities. All fitted functions are fixed before the audit. Predictions are strictly positive, and every baseline class probability is at least $\epsilon > 0$. Define $L = \log(1/\epsilon)$ and $a_0(H, U_0) = \mathrm{KL}(p_0(\cdot \mid H)\|q_0(\cdot \mid H, U_0))$. Then $0 \le a_0 \le L$. A witness is a measurable $b(H, U_0) \in [0, L]$, constructed without the audit outcome, satisfying

$$\Pr\{a_0(H, U_0) > b(H, U_0)\} \le \beta. \tag{9}$$

Theorem 4 (Failure-aware information certificate)**.** *Under the preceding assumptions,*

$$I(Y; G \mid H) \ge \mathbb{E}[D - b] - \beta L. \tag{10}$$

Let $\ell_\alpha(T_{1:n})$ *be any valid lower confidence bound on* $\mathbb{E}T$*, where* $T = D - b$*. Then*

$$I \ge \max\{0, \ell_\alpha(T_{1:n}) - \beta L\} \quad \textit{with probability at least } 1-\alpha. \tag{11}$$

Proof. On witness success, $a_0 \le b$; on failure, $a_0 - b \le L$. Thus $\mathbb{E}a_0 \le \mathbb{E}b + \beta L$. Proposition 1, applied with the independent inference randomizations included, gives Equation (10). Apply the stated lower confidence bound to $T = D - b$. For iid bounded scores, both betting and empirical Bernstein constructions are valid. □

The bound has three distinct sources of uncertainty: the approximation budget underlying b, the Monte Carlo failure allowance βL, and the sampling radius determined by n. More audit labels reduce only the last. More posterior draws may reduce the second and sharpen the witness, but do not remove a fixed inference bias.

We use betting on T for the natural-label and learned-representation audits, and empirical Bernstein comparisons (Maurer and Pontil, 2009; Waudby-Smith and Ramdas, 2024). With $a = -B - L$, a fixed mixture of products $\prod_i [1 - \eta + \eta(T_i - a)/(m - a)]$ tests $\mathbb{E}T \le m$; inversion gives ℓ_α. The supplement states the fractions and proof. Hoeffding is a valid, often looser alternative. For these constructions, assume iid audit records and $|D| \le B$. With M predeclared candidate audits, using α/M for each supplies simultaneous coverage by a union bound. Model or candidate selection using the same audit outcomes requires additional control; freezing the candidates beforehand avoids that issue.

For binary predictions, a convenient score bound is $q_1 = \sigma(\operatorname{logit} q_0 + v)$, $|v| \le B$, where σ is the logistic function. The Bernoulli log likelihood is 1-Lipschitz in its log odds, so $|D| \le B$. The bound applies to this constrained predictor, not to an unconstrained predictor that is only clipped after observing the results.

The witness-failure cost cannot simply be omitted. The inequality $A_0 - \mathbb{E}b \le \beta L$ is sharp under the witness assumptions; the supplement constructs an attaining example. More audit labels cannot remove this penalty. A smaller allowance requires control of failure severity.

Selecting before evaluation. A separate calibration set may select the score cap B and witness failure allowance β to maximize a predicted information bound. The supplement gives the objective. This is standard sample splitting, not a new selection theorem: conditional on calibration the procedure is frozen, so the independent audit remains valid. The experiment isolates whether selecting correction strength adds value beyond selecting witness confidence.

6 Experiments: what the procedure can and cannot establish

We test natural-label certification, categorical uncertainty, selection, and attribution. Complete protocols and the negative time-series case appear in the supplement.

Approximation improvement without new information. In a fixed two-component Gaussian mixture with constant inputs, correcting a missed rare component improves log score by .18400014 nats although $I = 0$. Its analytic reverse-KL budget is .01005 and the forward-error witness is .18400016: the corrected audit appropriately withholds an information claim. The supplement gives the optimization, quadrature and positive control.

Information certificates on natural outcomes. We split WDBC, Digits, and Wine into training, witness, and audit sets, using three fixed seeds per dataset (Wolberg et al., 1993; Alpaydin and Kaynak, 1998; Aeberhard and Forina, 1992). H is a binary context defined from one training-set summary; G contains all covariates. The baseline estimates outcome frequency within each context. The candidate averages 512 draws from a Laplace approximation to a continuous Bayesian logistic posterior, with its log-odds correction capped at two. No exact posterior normalizer is computed. Candidate approximation error need not be certified because it can only weaken the lower bound.

Separate natural labels give simultaneous binomial intervals for both context probabilities. Their endpoint forward KL gives a baseline-error witness with global failure allowance γ. An independent betting audit uses allowance α. Setting $\alpha = \gamma = .025/9$ controls the entire nine-analysis family at .05. This targets information beyond the coarse context under an iid population assumption; it does not condition on the full continuous covariate. Six bounds are positive (Table 2); Wine is inconclusive in every split. Within-context permutation tests reject independence in all nine analyses

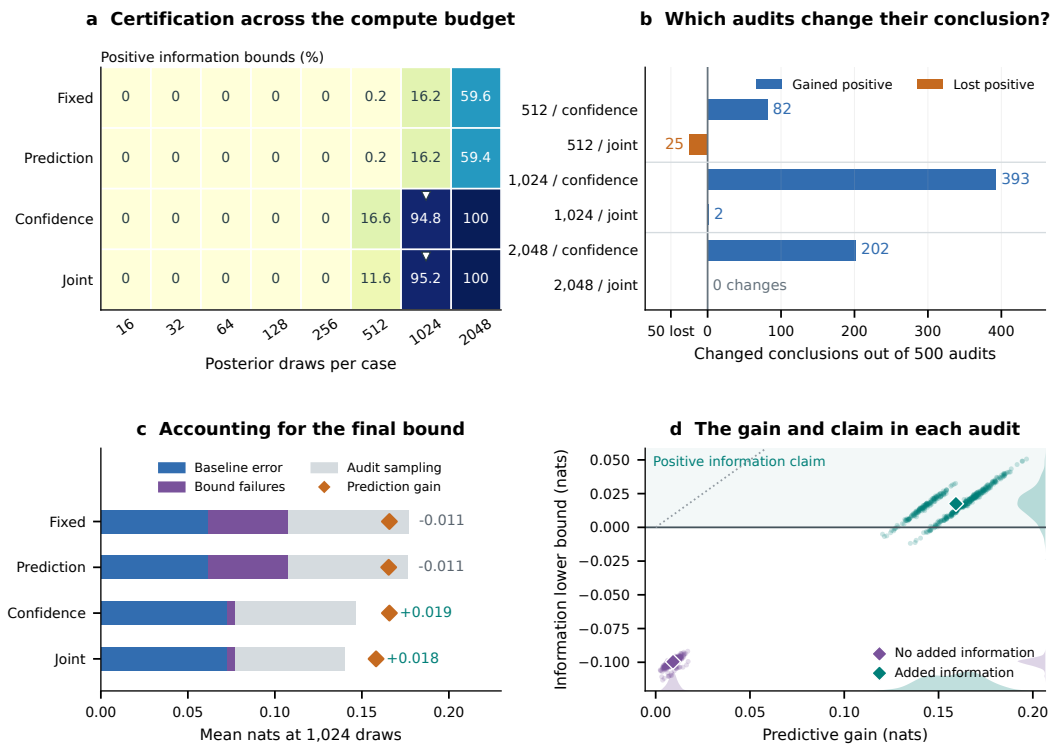

Figure 2: **Where certification improves, and why.** The fresh fitted-candidate experiment uses 300 calibration cases, 2,000 independent audit cases, and 500 repetitions per setting. (a) Color and numbers show positive-bound percentages across the full compute grid; white markers locate the first budget reaching 90%. (b) Paired counts of gained and lost positive conclusions. Each row labels draws / selection step: confidence compares confidence selection with fixed settings; joint compares joint with confidence selection. Unchanged conclusions are omitted. (c) At 1,024 draws, predictive gain (diamonds) must exceed all three stacked allowances; labels give the mean final bound. (d) Each point pairs a predictive gain with its information bound from the same repetition, at 1,024 draws with joint selection. Marginal densities show their distributions. The horizontal line separates positive from nonpositive information claims; the diagonal denotes zero total correction. Purple is the zero-information control; teal has $I = 0.1554$.

($p = .001$), illustrating the additional cost of quantification. The supplement reports every split, predictor, interval and score, as well as the separate enumerable Bayesian experiment that validates the posterior-to-predictive path under a specified model.

Table 2: Natural-label certificates with a fitted approximate Bayesian candidate. Three fixed splits per dataset; all nine bounds are simultaneous at 95% confidence under iid sampling. Gain and witness are split averages; the range reports the individual untruncated betting bounds, in nats. H is a training-defined binary context, not the full input.

Dataset	Audit n	Gain	Witness	Bound range	Positive
WDBC	172	0.378	0.118	0.111–0.149	3/3
Digits	540	0.360	0.028	0.220–0.259	3/3
Wine	54	0.264	0.279	-0.401—0.312	0/3

Selecting the information bound instead of the gain. The fresh experiment fits a candidate on 4,000 independent examples before the full compute grid is evaluated. A two-group population has conditional information .15540 nats from an added fair binary input. Known reverse-KL budgets and S independent Monte Carlo draws determine each witness; the supplement specifies the full generator.

We compare fixed $(B, \beta) = (2, 0.01)$, gain-based selection of B, bound-based selection of β with $B = 2$, and joint selection for the predicted information bound. At $S = 1024$, positive-certificate rates are 16.2%, 16.2%, 94.8%, and 95.2%; at $S = 2048$, they are 59.6%, 59.4%, 100%, and 100%. Confidence selection explains the improvement; joint selection adds little. Figure 2 reports the full range and paired changes. No information-bound violations occur in the fresh positive or null conditions. The earlier exploratory run remains separate in the supplement.

Known and sampled categorical baselines. The exact formula agrees with 48 partition enumerations and attains the reverse-KL budget and forward bound in 150 cases through $K = 100{,}000$. The new sampled-vector construction addresses a different problem: finite counts for small classes. Across 12,000 paired runs, it never equals the universal bound, whereas the earlier coordinate envelope is always vacuous at $K = 50$ and 100. Table 3 reports both bounds and actual error: the improvement is substantial but still leaves slack. The supplement gives all twelve settings, proofs, and numerical checks.

Table 3: Estimated categorical reports: 1,000 paired repetitions per row, reverse-KL budget .02, witness failure allowance .01. Means are in nats. The new bound is strictly below the universal bound in every run, but can remain loose. All twelve settings appear in the supplement.

K	Draws	Old bound	New bound	Actual error	Old vacuous
10	1000	2.906	0.416	0.029	98%
10	10000	0.335	0.160	0.025	0%
50	1000	5.505	1.514	0.060	100%
50	10000	5.175	0.501	0.038	100%
100	1000	6.748	2.424	0.090	100%
100	10000	6.034	0.764	0.047	100%

A learned representation: prediction is not information. A trained neural teacher generates labels on continuous 64-dimensional digit-image inputs. Both the raw-input probe and the representation probe are fitted on independent labels. When the added representation is a deterministic deeper layer, $I = 0$: prediction improves, but our corrected audit withholds an information claim. Adding an independent signal makes the teacher's $p_0(Y \mid H)$ computable by marginalizing that signal, supplying an exact witness. Table 4 compares the corrected audit with the holdout randomization test (HRT), using the true conditional feature sampler (Tansey et al., 2022). At 8,000 cases, betting detects the added signal in 98% of trials; HRT detects it in 100%. The latter tests independence, whereas our bound also quantifies information. All trained probes and repetitions are retained.

7 Related work and limitations

Usable information measures what a restricted predictor can extract; a deterministic feature can improve it without adding Shannon information (Xu et al., 2020; Ethayarajh et al., 2022). Variational bounds and Bayesian experimental design already target information (Barber and Agakov, 2003; Poole et al., 2019; Foster et al., 2019). Validated inference can supply approximation control (Huggins et al., 2020); our audit converts that control into the baseline-error allowance needed for attribution.

Conditional randomization and the HRT assume a valid law for $G \mid H$; conditional predictive impact also builds on valid conditional replacements (Candès et al., 2018; Tansey et al., 2022; Watson and Wright, 2021). Our audit instead needs control of $Y \mid H$, through baseline prediction error. These assumptions are complementary; neither method uniformly dominates.

Variable-importance inference is also close. LOCO measures changes in prediction, while algorithm-agnostic importance can target differences of optimal risks; with log loss, that population contrast is conditional information (Lei et al., 2018; Williamson et al., 2021; 2023). Decorrelated importance deliberately changes the target to address correlated inputs (Verdinelli and Wasserman, 2024). Floodgate and TV Floodgate already produce numerical lower bounds using a conditional feature law and a working predictor (Zhang and Janson, 2020; Wang et al., 2024). We therefore do

Table 4: **A learned representation can improve prediction without adding information.** A trained teacher generates labels on continuous 64-dimensional inputs. Both probes are fitted independently of the audit. Rates pool 50 repetitions for each of three fitted-probe seeds (150 per row). Prediction and corrected empirical Bernstein (EB) use fixed-sample bounds; betting uses the corrected score. HRT uses the exact conditional feature sampler. Its rejection tests independence; our bound also quantifies information.

Added input	Audit cases	Mean gain (nats)	Positive conclusion / rejection (%)			
			Prediction	Corrected EB	Betting	HRT
Deterministic layer	2000	0.0538	89.3	0.0	0.0	0.0
Deterministic layer	8000	0.0539	100.0	0.0	0.0	0.0
Layer + new signal	2000	0.0848	100.0	0.0	25.3	100.0
Layer + new signal	8000	0.0848	100.0	51.3	98.0	100.0

not claim that randomization-based methods only test independence. Our distinct input is an explicit finite-sample upper bound on baseline outcome-model error. A fresh paired TV Floodgate comparison in the supplement finds higher small-sample detection power for TV bounds, but a larger mean information bound from our audit at 8,000 cases.

Calibration also leaves information unresolved within groups sharing a score (Perez-Lebel et al., 2023). Decompositions for nested information sets formalize this distinction (Charpentier and Fernandes-Machado, 2026). Our added requirement is an upper baseline-error bound with a finite-sample decision rule; a decomposition or estimated grouping loss alone does not provide that guarantee.

The methodological contribution is the exact fixed-reference categorical envelope and its interval–simplex extension for sampled reports, coupled to failure-aware evaluation. Selection is an application of sample splitting; the experiments support confidence selection, not a general advantage for joint tuning. The log-score identity, divergence geometry, and information-bound optimization have established foundations (Gneiting and Raftery, 2007; Harremoës and Vajda, 2011; Sason and Verdú, 2016).

The hardest requirement remains a useful baseline-error bound. A small ELBO optimization gap or favorable chain diagnostic does not supply one automatically (Huggins et al., 2020; Vehtari et al., 2021). A bound valid under a Bayesian model does not establish that the model describes the data correctly. The exact categorical formula scales to large known vectors; uncertain high-dimensional vectors can still produce loose witnesses. The finite-sample theorem assumes independent cases and bounded scores. The selection experiment isolates a mechanism, not its prevalence in deployed systems. The natural-label study supplies an empirical witness only for a coarse, finite context. It does not solve scalable baseline validation under continuous high-dimensional conditioning. Shared-parameter exchangeability alone does not justify iid concentration; the supplement states conditional and independent-cluster extensions. These assumptions are essential, given known limits on general information estimation (McAllester and Stratos, 2020).

A nonpositive bound is inconclusive: candidate error, witness slack, weak information, or limited audit data may explain it. Useful deterministic representations remain useful for prediction (Xu et al., 2020).

8 Conclusion

A predictive gain supports an information claim only after baseline error and statistical uncertainty are controlled. The categorical conversion and failure accounting make that requirement explicit. Independent natural labels can supply the witness for a finite context; sampled multiclass reports require interval and normalization constraints. A learned-representation audit distinguishes information from improved probing. The audit connects proper-score evaluation, usable information, and validated inference (Gneiting and Raftery, 2007; Xu et al., 2020; Huggins et al., 2020).

AI USE STATEMENT

Generative AI was used to revise the manuscript language and prepare the submission files. The submission authors are responsible for reviewing the arguments, references, empirical records, and final disclosure before submission.

ETHICS STATEMENT

The controlled experiments use synthetic distributions. The enumerable Bayesian example uses model-based audit simulations; a separate study audits natural labels in three public datasets under an iid population assumption. The application uses public time-series benchmarks and makes no clinical or individual-level deployment claim. Its audit does not authorize any tested spectral deployment policy. A model-relative information guarantee should not be presented as evidence of real-world model correctness. The work concerns transparent reporting of probabilistic evidence and does not involve new human-subject data collection.

REPRODUCIBILITY STATEMENT

The accompanying package contains complete sources, proofs, executed experiment code, random seeds, machine-readable results, frozen application protocols, and saved predictions sufficient to replay the reported application gates and metrics. The controlled examples require no external data. Raw application archives and fitted classifier objects are not bundled; source manifests identify the archives used. The fast experimental path uses floating-point root calculations. A separate outward interval implementation checks the root conversion; it does not validate upstream approximation budgets or the entire numerical pipeline.

REFERENCES

Aeberhard, S.; and Forina, M. Wine [Dataset]. *UCI Machine Learning Repository, 1992. DOI: 10.24432/C5PC7J.* https://archive.ics.uci.edu/dataset/109/wine.

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

David Barber and Felix Agakov. The IM algorithm: A variational approach to information maximization. In *Advances in Neural Information Processing Systems*, volume 16, 2003. https://proceedings.neurips.cc/paper/2003/file/a6ea8471c120fe8cc35a2954c9b9c595-Paper.pdf.

Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135:1512–1519, 2009. https://doi.org/10.1002/qj.456.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B*, 80(3):551–577, 2018. https://doi.org/10.1111/rssb.12265.

Arthur Charpentier and Agathe Fernandes-Machado. Decomposing probabilistic scores: Reliability, information loss and uncertainty. *arXiv preprint arXiv:2603.15232*, 2026. https://arxiv.org/abs/2603.15232.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *ICML*, PMLR 162:5988–6008, 2022. https://proceedings.mlr.press/v162/ethayarajh22a.html.

Adam Foster, Martin Jankowiak, Eli Bingham, Paul Horsfall, Yee Whye Teh, Thomas Rainforth, and Noah Goodman. Variational Bayesian optimal experimental design. In *Advances in Neural Information Processing Systems*, volume 32, 2019. https://proceedings.neurips.cc/paper/2019/hash/d55cbf210f175f4a37916eafe6c04f0d-Abstract.html.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. https://doi.org/10.1198/016214506000001437.

Peter Harremoës and Igor Vajda. On pairs of f-divergences and their joint range. *IEEE Transactions on Information Theory*, 57(6):3230–3235, 2011. https://doi.org/10.1109/TIT.2011.2137353.

Jonathan Huggins, Mikolaj Kasprzak, Trevor Campbell, and Tamara Broderick. Validated variational inference via practical posterior error bounds. In *AISTATS*, PMLR 108:1792–1802, 2020. https://proceedings.mlr.press/v108/huggins20a.html.

Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R. J.; and Wasserman, L. Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association, 113(523):1094–1111, 2018.* https://doi.org/10.1080/01621459.2017.1307116.

Maurer, A., and Pontil, M. 2009. Empirical Bernstein Bounds and Sample Variance Penalization. *Proceedings of the 22nd Annual Conference on Learning Theory*. https://www.cs.mcgill.ca/~colt2009/papers/012.pdf.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *AISTATS*, PMLR 108:875–884, 2020. https://proceedings.mlr.press/v108/mcallester20a.html.

Alexandre Perez-Lebel, Marine Le Morvan, and Gaël Varoquaux. Beyond calibration: Estimating the grouping loss of modern neural networks. In *International Conference on Learning Representations*, 2023. https://openreview.net/pdf?id=6w1k-IixnL8.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *ICML*, PMLR 97:5171–5180, 2019. https://proceedings.mlr.press/v97/poole19a.html.

Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016. https://arxiv.org/abs/1508.00335.

Shah, R. D., and Peters, J. 2020. The Hardness of Conditional Independence Testing and the Generalised Covariance Measure. *The Annals of Statistics*, 48(3):1514–1538. https://arxiv.org/abs/1804.07203.

Wesley Tansey, Victor Veitch, Haoran Zhang, Raul Rabadan, and David M. Blei. The holdout randomization test for feature selection in black box models. *Journal of Computational and Graphical Statistics*, 31(1):151–162, 2022. https://doi.org/10.1080/10618600.2021.1923520.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, 16(2):667–718, 2021. https://doi.org/10.1214/20-BA1221.

Verdinelli, I.; and Wasserman, L. Decorrelated Variable Importance. *Journal of Machine Learning Research, 25(7):1–27, 2024.* https://www.jmlr.org/papers/v25/22-0801.html.

Wang, W.; Janson, L.; Lei, L.; and Ramdas, A. Total Variation Floodgate for Variable Importance Inference in Classification. *ICML, PMLR 235:50711–50725, 2024.* https://proceedings.mlr.press/v235/wang24ad.html.

David S. Watson and Marvin N. Wright. Testing conditional independence in supervised learning algorithms. *Machine Learning*, 110:2107–2129, 2021. https://doi.org/10.1007/s10994-021-06030-6.

Waudby-Smith, I., and Ramdas, A. 2024. Estimating Means of Bounded Random Variables by Betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27. https://academic.oup.com/jrsssb/article/86/1/1/7043257.

Williamson, B. D.; Gilbert, P. B.; Carone, M.; and Simon, N. Nonparametric variable importance assessment using machine learning techniques. *Biometrics, 77(1):9–22, 2021.* https://doi.org/10.1111/biom.13392.

Williamson, B. D.; Gilbert, P. B.; Simon, N.; and Carone, M. A General Framework for Inference on Algorithm-Agnostic Variable Importance. *Journal of the American Statistical Association, 118(543):1645–1658, 2023.* https://doi.org/10.1080/01621459.2021.2003200.

William Wolberg, Olvi Mangasarian, Nick Street, and W. Street. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository, 1993. https://doi.org/10.24432/C5DW2B.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020. https://openreview.net/forum?id=r1eBeyHFDH.

Zhang, L.; and Janson, L. Floodgate: Inference for Model-Free Variable Importance. *arXiv:2007.01283, 2020.* https://arxiv.org/abs/2007.01283.