AUDITING COMMON-THRESHOLD EXTRAPOLATION FOR RARE EVENTS IDENTIFIABILITY, PREDICTION, AND FAILURE MODES

Anonymous authors
Paper under double-blind review

ABSTRACT

When can common-threshold labels improve estimation of a rare-event probability? We develop a target-specific audit that distinguishes identifiability, statistical efficiency, and learning performance. For a locally identifiable target, the endpoint-to-common efficient variance ratio is $\mathcal{A} = (1-p)/(p\, g^\top I_c^+ g)$, where g is the target log-probability gradient and I_c is common-label Fisher information. A target direction outside the Fisher range rules out regular root-n estimation. Spectral diagnostics and a perturbation bound assess the stability of estimated information. In a two-rate benchmark, common-only extrapolation improves deep-tail accuracy, misspecification reverses the gain, and an added threshold restores identifiability. A prospective neural study evaluates five targets at three model sizes with approximately matched parameter counts. A pointwise Fisher score predicts the mean direction of the error difference in all 12 cells outside its indifference band, with rank correlation 0.917 across all 15 cells. At the deepest target, endpoint-to-common error ratios are 1.67–1.74. An independent 20,000-observation common-label pilot identifies harm at three targets and abstains at two, falling short of its prespecified decision-coverage criterion. Temporal tests show mixed effects on USGS earthquakes and 3.5–16.8% lower held-out log loss in four gas-turbine target–year cells. The resulting approach connects label design to testable learning predictions through target identifiability and information margins.

1 INTRODUCTION

Suppose a damaging event has probability $p = 10^{-6}$. Estimating it accurately in relative terms requires many endpoint observations. More frequent labels at lower thresholds may also be informative. We ask: *at the same sample size, when can common labels alone estimate a rare target more precisely than its endpoint indicator?* Multi-task learning encourages sharing across related labels (Caruana, 1997), and survival models jointly represent multiple time points or risks (Yu et al., 2011; Lee et al., 2018). Our question concerns the information in the labels before choosing a learning architecture.

Whether tasks benefit from shared learning depends on data, capacity, and optimization (Standley et al., 2020). Gradient conflict can impede joint optimization (Yu et al., 2020), while asymmetric task relations motivate selective transfer (Lee et al., 2016). Learning theory characterizes the benefits of shared representations in terms of task count, sample size, and representation specificity (Maurer et al., 2016). Transferability scores predict downstream accuracy from labeled target examples (Nguyen et al., 2020); high-dimensional analyses identify positive and negative transfer regimes (Yang et al., 2025); and capacity–redundancy analyses examine interference under shared representations (Khan, 2026). These approaches motivate a target-specific assessment of when sharing can help.

We compare two observation schemes: one retains the endpoint indicator, and the other retains only common-threshold labels. The relevant quantity is the information each provides about the target probability. Fisher matrices have been used to compare whole tasks (Achille et al., 2019), and Fisher-information ratios motivate active querying (Sourati et al., 2017). For a target with gradient g, minimizing $g^\top I^+ g$ is the classical c-optimal design criterion (Elfving, 1952). We use this foundation to normalize efficiency against endpoint-only estimation, identify target directions missing from the common labels, and assess uncertainty in the resulting prediction. Rare-event methods also gain

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.

efficiency from structural assumptions, including tail laws (Pickands, 1975; Embrechts et al., 1997) and Markov structure (Cheng & Weare, 2024). Our audit makes the role of such assumptions explicit in common-label extrapolation.

Common Event Tethering (CET) (Lanners et al., 2024) is a closely related approach that regularizes multi-label predictors toward a related common event and evaluates two clinical applications. We include a neural CET adaptation as a full-label baseline. The common-only comparison also permits an earlier diagnostic: whether the observed labels identify the rare target at all. This question differs from estimating hitting-time or "A before B" queries of an already fitted process (Boyd et al., 2023); here, we study which labels identify a rare query and how precisely they can estimate it.

Our contributions are:

1. We specialize local asymptotic theory and c-optimal design to common-threshold labels, deriving a necessary Fisher-range condition, an endpoint-relative efficiency ratio, and its rare-tail limit.
2. We provide a spectral diagnostic and a deterministic perturbation bound for estimated Fisher information. With valid uncertainty bounds, these support a three-way decision: help, harm, or abstain.
3. We use a two-rate competing-risk benchmark to separate identifiability, efficiency, and misspecification. An independent replication tests the deep-tail gains, and explicit failure cases show when those gains disappear.
4. We prospectively test a pointwise Fisher score in a 15-cell neural study with approximately matched parameter counts. A separate common-label pilot measures decision uncertainty; a CET baseline and two public temporal datasets test the comparison in additional settings.

2 Common-only estimation

Let $Y_a \in \{0,1\}$ indicate a rare target event with probability $p_a(\theta) \in (0,1)$ under a regular parametric model P_θ, $\theta \in \mathbb{R}^d$. An endpoint-only estimator observes $Y_a^{(1)}, \dots, Y_a^{(n)}$. A common-only estimator observes a categorical summary C of labels at thresholds below a, without access to the target indicator. Write

$$g_a(\theta) = \nabla_\theta \log p_a(\theta), \qquad I_{\mathrm{c}}(\theta) = \mathbb{E}_\theta[s_C(C;\theta) s_C(C;\theta)^\top].$$

Here $s_C(c;\theta) = \nabla_\theta \log P_\theta(C = c)$, and I_{c}^+ denotes the Moore–Penrose pseudoinverse. Information is computed from the joint categorical likelihood, which preserves the dependence among nested labels.

Scope of the comparison. The ratio $\mathcal{A}$ compares two label sets at the same sample size n. Observing both sets defines a full-label experiment with label $F = (C, Y_a)$. Each coarsened score is the conditional expectation of the full score, so Fisher-information data processing implies that, under correct specification and target identifiability, an efficient full-label estimator has asymptotic variance no larger than either coarsened estimator. Practical joint learning can nevertheless suffer from misspecification, regularization, or optimization. Unequal labeling costs are outside this equal-n comparison.

2.1 Efficiency and identifiability

Theorem 1 (Endpoint versus common-only efficiency)**.** *At an interior θ, assume differentiability in quadratic mean, locally constant rank of I_{c}, and an efficient regular estimator of the identifiable coordinates from C. Suppose that $\log p_a$ is locally a smooth function of those coordinates, with $g_a \in \mathrm{Range}(I_{\mathrm{c}})$ and $g_a^\top I_{\mathrm{c}}^+ g_a > 0$. Then*

$$\sqrt{n}\{\log \widehat{p}_{\mathrm{c}} - \log p_a\} \Rightarrow N\big(0, g_a^\top I_{\mathrm{c}}^+ g_a\big)\,.$$

For the endpoint MLE $\widehat{p}_{\mathrm{e}} = n^{-1}\sum_i Y_a^{(i)}$,

$$\sqrt{n}\{\log \widehat{p}_{\mathrm{e}} - \log p_a\} \Rightarrow N\left(0, \frac{1-p_a}{p_a}\right).$$

Hence the endpoint-to-common asymptotic variance ratio is

$$\mathcal{A}_a(\theta) = \frac{1 - p_a}{p_a \, g_a^\top I_c^+ g_a}. \tag{1}$$

If $g_a \notin \mathrm{Range}(I_c)$, no regular estimator based only on C can estimate $\log p_a$ at the root-n rate.

Thus $\mathcal{A}_a > 1$ favors common-only estimation, while $\mathcal{A}_a < 1$ favors endpoint-only estimation. The proof in Appendix A follows from local asymptotic efficiency and the delta method (van der Vaart, 1998). The range condition is necessary: a pseudoinverse cannot recover a missing target direction. Sufficiency also requires the theorem's local target-identifiability assumption.

The variance ratio also determines an absolute-error ratio when the scaled absolute log errors are uniformly integrable. In that case, the normal limits give $\mathbb{E}|\log \widehat{p}_e - \log p|/\mathbb{E}|\log \widehat{p}_c - \log p| \to \sqrt{\mathcal{A}}$. This expectation statement requires smoothing the endpoint count, for example as $(K+1/2)/(n+1)$, because the unsmoothed MLE is zero with positive probability.

Corollary 1 (Rare-tail gain)**.** *Under the identifiable-target conditions of Theorem 1, if $p_a \to 0$ and $p_a g_a^\top I_c^+ g_a \to 0$, then $\mathcal{A}_a \to \infty$.*

The large-n limit is taken at each fixed target before taking the tail limit. A rarer target therefore offers increasing gains only if the common labels retain its direction and its log-probability variance grows slowly enough.

Proposition 1 (Ridge-stable range stress test)**.** *Let $I \succeq 0$, let P_0 project onto $\mathrm{Null}(I)$, and define $V_\lambda(g) = g^\top (I + \lambda I_d)^{-1} g$ for $\lambda > 0$. Then*

$$V_\lambda(g) = \sum_{\mu_j > 0} \frac{(u_j^\top g)^2}{\mu_j + \lambda} + \frac{\|P_0 g\|^2}{\lambda}.$$

Thus $V_\lambda(g) \to \infty$ when the target has a null-space component. If $g \in \mathrm{Range}(I)$ and μ_+ is the smallest positive eigenvalue, then

$$0 \le g^\top I^+ g - V_\lambda(g) \le \lambda \|g\|^2 / \mu_+^2.$$

Proposition 2 (Plug-in Fisher perturbation certificate)**.** *Let $I, \widehat{I} \succeq 0$, $g, \widehat{g} \in \mathbb{R}^d$, and suppose $\|\widehat{I} - I\|_{\mathrm{op}} \le \epsilon_I$ and $\|\widehat{g} - g\| \le \epsilon_g$. For $\lambda > 0$, write $V_\lambda = g^\top (I + \lambda I_d)^{-1} g$ and define $\widehat{V}_\lambda$ analogously. If $G \ge \|g\|$, then*

$$|\widehat{V}_\lambda - V_\lambda| \le \Delta_\lambda := \frac{2G\epsilon_g + \epsilon_g^2}{\lambda} + \frac{G^2 \epsilon_I}{\lambda^2}. \tag{2}$$

If additionally $I \succ 0$ with smallest eigenvalue μ, then for $V_0 = g^\top I^{-1} g$, $|\widehat{V}_\lambda - V_0| \le \Delta_\lambda + \lambda G^2/\mu^2$.

Proposition 1 provides a stable way to examine weakly identified directions. For each cutoff τ, let I_τ retain the eigenvalues of I above τ, and let P_τ project onto their eigenspace. Report the retained variance $g^\top I_\tau^+ g$ and the normalized discarded component $\rho_\tau = \|(I_d - P_\tau) g\|/\|g\|$ across a grid of cutoffs. A diverging ridge curve or a substantial discarded component indicates that the conclusion depends on poorly identified directions. For estimated information, repeating the sweep across bootstrap fits also reveals sensitivity to estimation error.

Proposition 2 separates gradient error, information error, and ridge bias. An observable bound on the gradient norm is $G = \|\widehat{g}\| + \epsilon_g$. If $B = (1-p)/p$ lies in $[B_L, B_U]$ and the bounds above give $0 < V_L \le V_0 \le V_U$, then $\mathcal{A} \in [B_L/V_U, B_U/V_L]$. A ratio interval entirely above one certifies help; an interval entirely below one certifies harm; otherwise, the rule abstains. Concentration inequalities or bootstrap methods can supply uncertainty bounds. The certificate is conditional on their validity; in particular, it does not make bootstrap coverage exact.

Operational audit. For a proposed set of common labels, the audit has five steps: (i) specify the joint categorical likelihood; (ii) check the Fisher-range condition and spectral stability; (iii) estimate $\mathcal{A}_a$ with uncertainty and require a practically meaningful margin above one; (iv) vary nuisance parameters and plausible forms of misspecification; and (v) validate the prediction on an independent random stream or data split chosen in advance. For a finite pool of independently allocated label

designs with information matrices I_j, one can select thresholds through the regularized c-optimal program $\min_{w\in\Delta} g^\top(\sum_j w_j I_j + \lambda I_d)^{-1} g$, followed by a pre-specified rounding rule. An exact range violation rules out regular common-only estimation; numerical instability indicates a need for more information or abstention.

3 A TWO-RATE COMPETING-RISK BENCHMARK

The following model lets us study identifiability and efficiency explicitly. With probability $1/2$, an observation belongs to a competing lower event. Otherwise, its upper mark M has piecewise exponential survival

$$S(m) = \exp\{-r_{\mathrm{p}} \min(m,t) - r_{\mathrm{q}}(m-t)_+\}, \quad t = 2.$$

The common thresholds are $c_1 = 0.5$ and $c_2 = 2.5$; the target is $a > c_2$. One observation falls into five categories with probabilities

$$q(\theta, a) = \tfrac{1}{2}(1,\ 1 - S(c_1),\ S(c_1) - S(c_2),\\ S(c_2) - S(a),\ S(a)).$$

The endpoint count is the fifth category. The common-only moment estimator merges the last two categories, concealing the target count. Write $(N_0, \ldots, N_4)$ for counts in the displayed order. The implementation clips the empirical survival probabilities $1 - 2N_1/n$ and $2(N_3 + N_4)/n$ to $[10^{-15}, 1 - 10^{-15}]$. When no clipping is needed, the estimates are

$$\widehat{r}_{\mathrm{p}} = -\frac{\log(1 - 2N_1/n)}{c_1}, \qquad \widehat{r}_{\mathrm{q}} = \left[\frac{-\log\{2(N_3 + N_4)/n\} - t\widehat{r}_{\mathrm{p}}}{c_2 - t}\right]_+,$$

and $\widehat{p}_{\mathrm{c}} = \frac{1}{2} S(a; \widehat{r}_{\mathrm{p}}, \widehat{r}_{\mathrm{q}})$. The true model sets $r_{\mathrm{p}} = r_{\mathrm{q}} = \exp(0.25h)$ for contexts $h \in \{-1, 1\}$.

The efficient-information crossover $\mathcal{A} = 1$ occurs at $p_\times = 2.17 \times 10^{-4}$ for $h = -1$ and 3.15×10^{-4} for $h = 1$. We test a pre-tail point $p_{\mathrm{pre}} = 4.34 \times 10^{-4}$ and three tail points $p_1 = 1.09 \times 10^{-4}$, $p_2 = 2.17 \times 10^{-5}$, and $p_3 = 2.17 \times 10^{-6}$. The efficient variance ratios at p_1, p_2, p_3 are $(1.61, 5.21, 31.86)$ for $h = -1$ and $(1.92, 5.81, 33.50)$ for $h = 1$. The moment estimator need not attain these ratios; its asymptotic variance is derived separately in Appendix B.1.

4 EXPERIMENTS

Protocol. The original experiment pairs endpoint-only and common-only estimates from the same multinomial count vector across 200 seeds, with $N_{\mathrm{hi}} = 920{,}987{,}204$ observations. Count sampling is distribution-equivalent to simulating all episodes. We report the ratio of seed-mean absolute log errors, $R = \overline{E}_{\mathrm{endpoint}}/\overline{E}_{\mathrm{common}}$. A paired delta-method standard error for $\log R$ gives a one-sided 95% lower confidence bound (LCB). The primary criterion requires this bound to exceed one in all six context–tail cells. To measure the trend with rarity, we regress $\log R$ on $\log_{10}(1/p)$ and propagate the pointwise variances. An independent replication uses a new root seed and seed IDs 5000–5199, specified before sampling. A lossless 3,200-row replay reproduces every saved aggregate to 10^{-12}; Appendix B gives the protocols and evidence hashes.

Table 1: Primary high-N results. $R > 1$ favors common-only extrapolation; LCB is a one-sided 95% lower bound. The original study passes 6/6 cells and the independent replication passes 5/6, with the first row falling below the criterion.

h	p	Original R	LCB	Independent R	LCB
-1	1.09e-04	1.15	1.01	0.99	0.88
-1	2.17e-05	1.93	1.71	1.91	1.66
-1	2.17e-06	4.80	4.25	4.16	3.65
1	1.09e-04	1.16	1.02	1.18	1.04
1	2.17e-05	1.97	1.75	1.92	1.70
1	2.17e-06	4.92	4.35	4.99	4.44

4.1 THE DEEP-TAIL TREND REPLICATES; THE BOUNDARY CELL DOES NOT

Common-only estimation becomes more effective deeper in the tail (Table 1). All six original cells pass the primary criterion. In the independent replication, ratios at p_2 and p_3 range from 1.91 to 4.99, with all lower bounds above 1.66. The rarity slopes are also positive: 0.841 (LCB 0.734) for $h = -1$ and 0.858 (LCB 0.759) for $h = 1$. The boundary cell $h = -1, p_1$ gives $R = 0.99$ with lower bound 0.88, so the replication fails the pre-specified six-cell conjunction. The results support a reproducible deep-tail gain, with uncertainty near the crossover.

A post hoc sensitivity analysis uses 20,000 paired nonparametric bootstrap resamples under a protocol fixed before resampling. Its lower bounds differ from the primary delta-method bounds by at most 0.006 and preserve every cell-level decision, including the failed replication cell (Appendix B).

Figure 1 compares R^2 with the efficient variance ratio $\mathcal{A}$ and the moment estimator's asymptotic ratio $\mathcal{A}_{\text{mom}}$. The empirical ratios follow the increasing tail trend. Their departure from $\mathcal{A}$ reflects both moment-estimator inefficiency and sampling variability: mean $|\log(R^2/\mathcal{A})|$ across the six cells is 0.32 in the original study and 0.42 in the replication (maximum 0.61). A separate 500-replication study varies the budget from 10 to 2,000 expected endpoint positives Np, using a Jeffreys-smoothed endpoint estimate to handle zero counts. At the rarest target, common-only estimation reduces mean absolute log error by factors of 4.4–5.5 throughout this range.

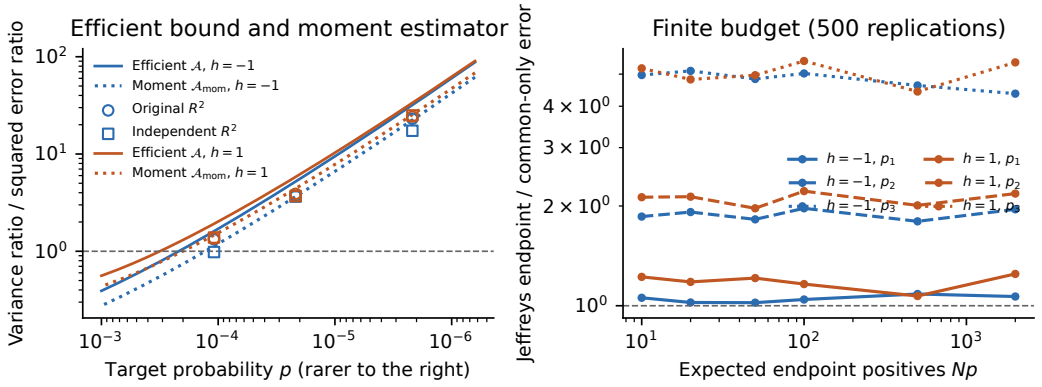

Figure 1: Left: efficient variance ratio $\mathcal{A}$ (solid), moment-estimator ratio $\mathcal{A}_{\text{mom}}$ (dotted), and squared empirical error ratios R^2 (markers). Right: finite-budget ratios against a Jeffreys-smoothed endpoint; line styles distinguish the three tail probabilities. Horizontal lines mark equality.

4.2 A DETERMINISTIC NO-GAIN ABLATION

Remove c_2 and retain only $c_1 < t$. The common Fisher information then has rank one, with range $\text{span}(1, 0)$, while $g_a = (-t, -(a - t))$ has a nonzero post-knot component. Theorem 1 therefore rules out regular common-only estimation. If we instead retain the endpoint, can the remaining c_1 label improve its estimate? To isolate this comparison, we discard the competing-event identity, leaving probabilities $(1 - q, q - p, p)$, where q is the probability of exceeding c_1 and p is the target probability. In this saturated nested-label model, the categorical MLE satisfies

$$\widehat{p}_{\text{joint}} = N_{\text{target}}/n = \widehat{p}_{\text{e}}$$

for every nested-label count vector. A constrained two-rate fit has the same identity whenever the empirical proportions lie in its feasible interior; boundary-constrained fits need not. The saturated model thus illustrates a second limitation: the common label provides no gain even when the endpoint is observed.

4.3 PROSPECTIVE PREDICTION OF LEARNED BENEFIT AND HARM

Does information in the labels predict the performance of a trained neural network? We test this question with context $x \sim \text{Unif}[-1, 1]$ and hazards $\log r_{\text{p}}(x) = -0.10 + 0.35x$ and $\log r_{\text{q}}(x) = 0.05 - 0.25x$. An initial 30-seed study with width-16 networks found no significant common-only

improvement under its primary paired criterion, while secondary full-label contrasts were positive (Appendix D). That study used unequal parameter counts and motivated the prospective comparison below.

We specified five targets and three model sizes before running study seeds 8000–8023. Each seed provides 20,000 training and 5,000 validation observations. Endpoint-only and two-hazard common-only networks use the same optimization procedure and approximately matched parameter counts: widths $(8, 6)$, $(16, 12)$, and $(32, 24)$ give $(25, 26)$, $(49, 50)$, and $(97, 98)$ parameters, respectively. Before training, we compute the pointwise Fisher score

$$Q_a = \frac{\langle \sqrt{(1 - p_a(x))/p_a(x)} \rangle_x}{\langle \sqrt{g_a(x)^\top I_c(x)^{-1} g_a(x)} \rangle_x}, \tag{3}$$

where $\langle \cdot \rangle_x$ denotes averaging over a fixed 2,001-point grid on $[-1, 1]$. The information matrix $I_c(x)$ and gradient $g_a(x)$ use the two log-hazard coordinates. The score Q_a is a ratio of averaged pointwise standard deviations; $\mathcal{A}_a$ is an efficient variance ratio. We test whether Q_a predicts performance when a network shares parameters across contexts. The observed error ratio R_a divides the endpoint-only seed-mean error by the common-only seed-mean error, with each error defined as $E_a = \langle |\log \widehat{p}_a(x) - \log p_a(x)| \rangle_x$. Using the pre-specified indifference band $[0.95, 1.05]$, the score predicts harm at $a = 4, 5, 6$, makes no directional prediction at $a = 7$, and predicts benefit at $a = 8$.

Table 2: Prospective validation with approximately matched parameter counts. Q_a is the pointwise Fisher score. Small, medium, and large report observed error ratios R_a; values above one favor common-only estimation. CET reports the endpoint-to-CET error ratio for a 50-parameter full-label baseline.

a	Q_a	Small	Medium	Large	CET
4	0.53	0.48	0.46	0.50	0.65
5	0.58	0.45	0.48	0.47	0.74
6	0.73	0.54	0.52	0.55	0.77
7	1.00	0.83	0.79	0.78	1.03
8	1.43	1.67	1.69	1.74	0.97

The score meets the pre-specified prediction criterion: at least 10/12 correct directions and Spearman correlation of at least 0.75. All 12 cells outside the indifference band have the predicted mean direction, and the correlation between $\log Q_a$ and $\log R_a$ across all 15 cells is 0.917. At $a = 8$, R_a ranges from 1.67 to 1.74; at $a = 4, 5, 6$, it ranges from 0.45 to 0.55. At $a = 7$, where the score makes no directional prediction, the endpoint performs better ($R_a = 0.78$–0.83). Only the large-model $a = 8$ cell has a positive one-sided 95% bootstrap lower bound for the paired endpoint-minus-common error difference. Thus the cross-cell criterion establishes directional prediction, while evidence for individual gains is more limited. Cells share targets and training data and are dependent. Integrated Bernoulli log-score regret has the same direction in the 12 cells with a directional prediction. Brier regret favors the endpoint even at $a = 8$, illustrating the difference between squared probability error and relative tail accuracy.

Our neural CET adaptation (Lanners et al., 2024) uses a shared encoder, rare- and common-event Bernoulli heads, and an L2 penalty tethering the heads, with its strength selected on validation data. The endpoint-to-CET error ratios are 0.65, 0.74, 0.77, 1.03, and 0.97 at $a = 4, \ldots, 8$; no paired endpoint-minus-CET error-difference lower bound exceeds zero. CET observes both label sets, so the common-only score does not directly predict its performance. Under this implementation and tuning grid, it does not reproduce the structured common-only model's deep-tail gain (Figure 2).

4.4 A COMMON-LABEL PILOT AUDIT WITH ABSTENTION

Can the same predictions be made when the data-generating law is unknown? We evaluate this with an independent 20,000-observation pilot, under a protocol fixed before sampling. The pilot contains only contexts and four-category common labels. We fit four log-linear hazard coefficients by the exact categorical likelihood and use them to compute Q_a. Each of 1,000 nonparametric bootstrap resamples refits the coefficients and recomputes the pointwise information matrices. These intervals estimate uncertainty in Q_a directly; the deterministic perturbation bound provides a separate route to certification.

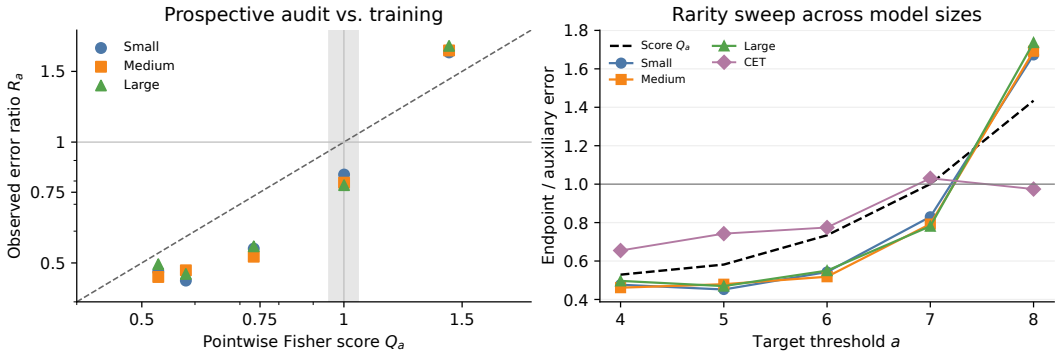

Figure 2: Prospective neural validation across all 15 target–model cells. Left: pointwise Fisher score Q_a versus observed integrated-error ratio R_a; the diagonal denotes numerical agreement, while the lines at one separate benefit from harm; shading marks the score's indifference band. Right: the three model sizes and the neural CET adaptation across the five targets. All plotted error ratios compare seed means.

For $a = 4, \ldots, 8$, the estimated scores and nominal Bonferroni simultaneous 95% percentile intervals are 0.48 [0.41,0.56], 0.49 [0.39,0.64], 0.58 [0.42,0.83], 0.74 [0.48,1.17], and 0.99 [0.59,1.76]. Combined with the pre-specified practical band $[0.95, 1.05]$, these yield harm at $a = 4, 5, 6$ and abstention at $a = 7, 8$. The three harm decisions agree with the observed region of poorer common-only performance, but the pilot leaves the oracle score's deep-tail benefit unresolved.

The decision-coverage criterion requires at least ten correct cells, zero wrong decisions, an abstention target, and both help and harm decisions. Repeating the five target-level calls across three model sizes gives nine correctly classified cells and six abstaining cells. The nine correct cells reflect three harm decisions. The pilot therefore fails both the minimum-count and the help/harm requirements.

5 Assumption and External-Data Stress Tests

Tail curvature. We generate from post-knot cumulative hazard $H(t + z) = rt + rz + r\kappa z^2/(2L)$ with $L = 20$, restricted to the queried range $0 \le z \le a - t$, where $1 + \kappa z/L > 0$. We continue to fit the piecewise-constant model. At p_3, $\kappa = \pm 0.25$ induces mean signed log bias of magnitude 0.43–0.90 and reduces endpoint-to-common error ratios to 0.020–0.041 (Figure 3, left). The correctly specified $\kappa = 0$ ratio is 4.73–5.44. Here, extrapolation bias overwhelms the variance reduction. The efficiency ratio in Eq. 1 depends on a correctly specified tail family.

Proposition 3 (Information completion under curvature)**.** *Parameterize the post-knot cumulative hazard by $H(t + z) = r_{\mathrm{p}}t + r_{\mathrm{q}}z + cz^2/2$. For $a > c_2 > t$, labels at $c_1 < t < c_2$ do not identify the target gradient. Adding any distinct $c_3 > c_2$ below a makes the three-parameter common experiment locally full rank, provided its category probabilities are interior.*

We set $c_3 = 4.0$ and specified the curvature grid, 200 replications, and new random streams before fitting the three-threshold model. At p_3, its largest absolute signed log bias is 0.0036, versus 1.80 for the misspecified two-threshold fit. Its endpoint-to-common error ratio ranges from 0.26 to 2.28 and exceeds one in only 6/10 cells. Both curves in Figure 3 use the same paired draws. The extra threshold restores identifiability, but the information it provides is sufficient to outperform the endpoint in only part of the grid.

USGS temporal transfer. To test the comparison on real temporal data, we use reviewed earthquakes in a California bounding box with magnitude $M \ge 3$ from 2000–2025 (U.S. Geological Survey, 2026). For each test year 2010–2025, the preceding ten years form the training fold. We compare a target-specific Jeffreys estimate against a one-rate exponential tail fitted only to common labels $M \ge 3.5$ and $M \ge 4.0$, motivated by the Gutenberg–Richter relation (Gutenberg & Richter, 1944). An exponential MLE fitted to exact magnitudes provides a benchmark with richer observations under the same one-rate family. We evaluate held-out Bernoulli log loss.

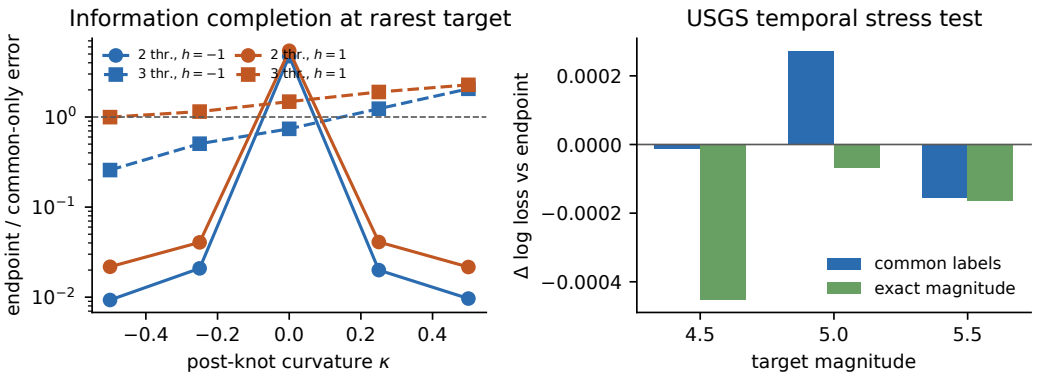

Figure 3: Left: curvature reverses the two-threshold gain; a third common threshold nearly removes empirical extrapolation bias but is not uniformly more efficient than the endpoint. Right: event-weighted USGS log-loss differences from the endpoint-only estimate; lower is better.

The filtered catalog contains 11,991 events; the 16 held-out test years contain 8,219 events. Event-weighted endpoint log losses are 0.1256, 0.0434, and 0.0202 at $M \geq 4.5, 5.0, 5.5$. Relative changes for common-only estimation are -0.01%, $+0.63\%$, and -0.77%; common-only estimation improves on the endpoint in 6, 6, and 7 of 16 folds, respectively. Exact magnitudes improve the event-weighted aggregate by 0.36%, 0.16%, and 0.82%, most consistently at $M \geq 5.5$ (11/16 folds). The gains are small and vary across years. Overlapping training windows, earthquake clustering, heterogeneous magnitude types, and catalog completeness make this a descriptive assessment of temporal transfer.

Gas-turbine temporal transfer. The UCI gas-turbine dataset contains 36,733 hourly sensor aggregates from an operating power plant in 2011–2015 (UCI Machine Learning Repository, 2019). Following its recommended temporal protocol, we fit on 22,191 observations from 2011–2013 and test separately on 2014 and 2015. The endpoint is a target-specific Jeffreys estimate for NOx concentration above 100 or 110 mg/m^3. The common-only working model sees only four bins cut at 70, 80, and 90 mg/m^3 and extrapolates an exponential excess tail. Its fitted efficient variance ratios $\mathcal{A}$ are 1.91 and 3.34. Table 3 reports all four target–year cells: common-only log loss falls by 3.5–16.8%, with the same direction under Brier score. Designed after inspection of the public data, this study provides a descriptive comparison under temporal shift in a second physical domain.

Table 3: UCI gas-turbine held-out Bernoulli log loss. The efficient variance ratio $\mathcal{A}$ is fitted on 2011–2013 data; negative relative change favors common-only.

a	Test year	$\mathcal{A}$	Endpoint	Common	Rel. change
100	2014	1.91	0.0439	0.0403	-8.2%
100	2015	1.91	0.0590	0.0568	-3.9%
110	2014	3.34	0.0146	0.0122	-16.8%
110	2015	3.34	0.0282	0.0272	-3.5%

6 Discussion and Limitations

The audit connects common-label design to information about a specific rare target. It complements survival networks that model event-time distributions and competing risks (Lee et al., 2018; Kvamme et al., 2019; Rindt et al., 2022) by asking, before training, whether the available labels retain the required target direction. The experiments then test whether this information advantage survives estimation and learning. Performance depends on both the learner's efficiency and the accuracy of its tail model. For first-passage applications, Appendix F gives an exact verification framework: running upper and lower records preserve finite-horizon competing crossing queries and the Fisher information in the records.

The theory is local and regular, so finite samples, nuisance estimation, boundary constraints, and misspecification can dominate its predictions. The perturbation bound quantifies the effect of information-estimation error; obtaining useful bounds in high dimensions remains a separate challenge. The neural study uses one context variable, one synthetic family, and a correctly structured two-hazard head. Its score Q_a aggregates pointwise information rather than computing the Fisher matrix of the shared network. The pilot estimates four hazard coefficients to evaluate the same score, with approximate bootstrap intervals. Its abstention at the two deepest targets shows the uncertainty that remains even in this setting. Broader architectures and survival-model comparisons would test the reach of the observed predictive relation.

The external studies are marginal tail forecasts without covariates. The USGS comparison is affected by overlapping training windows, event clustering, and heterogeneous magnitude types. The UCI study was designed after data inspection and includes only two held-out years. Extending the audit to conditional forecasts, wider temporal evaluations, and unequal labeling costs would address important aspects of practical label selection.

7 Conclusion

Common-threshold labels can improve rare-event estimation when they identify the target direction and provide enough information along it. The benchmark shows both reproducible deep-tail gains and two distinct failure modes: missing information and extrapolation bias. In the prospective neural study, the pointwise Fisher score predicts the mean direction of benefit and harm across three model sizes, while the estimated pilot score identifies harm and abstains in the deep tail. The CET comparison and temporal studies further show that realized gains depend on the method and distribution. Together, range checks, uncertainty estimates, and independent validation provide a testable approach to designing common labels for rare targets.

AI use statement

Generative AI tools were used to polish the language of this manuscript. The authors are responsible for the final content.

Ethics statement

The synthetic experiments involve no human subjects. The external experiments use the public USGS earthquake catalog and a CC BY 4.0 industrial-emissions dataset; neither supports an operational hazard, maintenance, or safety claim.

Reproducibility statement

Code and reproducibility materials are available anonymously at `https://anonymous.4open.science/r/rare-event-audit-artifact-8351/`, including all study protocols, evidence hashes, seed registries, lossless count and training rows, aggregate results, and figure and table sources. Scripts recompute the experiments and regenerate figures and tables from saved results. Appendices A, B, and F give the proofs, estimation and inference details, and first-passage oracle tests. These materials also record the USGS query, UCI archive, selection rules, hashes, and temporal splits.

References

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C. Fowlkes, Stefano Soatto, and Pietro Perona. Task2Vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6430–6439, 2019.

Alex Boyd, Yuxin Chang, Stephan Mandt, and Padhraic Smyth. Probabilistic querying of continuous-time event sequences. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 10235–10251. PMLR, 2023.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997. doi: 10.1023/A:1007379606734.

Yichen Kelly Chen, Sören Dittmer, Kinga Bernatowicz, Josep Arús-Pous, Kamen Bliznashki, John Aston, James H. F. Rudd, Carola-Bibiane Schönlieb, James Jones, and Michael Roberts. SurvSurf: A partially monotonic neural network for first-hitting time prediction of intermittently observed discrete and continuous sequential events. arXiv:2504.04997, 2025. URL https://arxiv.org/abs/2504.04997.

Xiaoou Cheng and Jonathan Weare. The surprising efficiency of temporal difference learning for rare event prediction. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Gustav Elfving. Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics*, 23(2):255–262, 1952. doi: 10.1214/aoms/1177729442.

Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer, 1997.

Beno Gutenberg and Charles F. Richter. Frequency of earthquakes in California. *Bulletin of the Seismological Society of America*, 34(4):185–188, 1944.

Asif Khan. Capacity and redundancy trade-offs in multi-task learning. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337, pp. 2934–2957. PMLR, 2026.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019.

Quinn Lanners, Qin Weng, Marie-Louise Meng, and Matthew M. Engelhard. Common event tethering to improve prediction of rare clinical events. In *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence*, volume 244, pp. 2136–2162. PMLR, 2024.

Changhee Lee, William R. Zame, Jinsung Yoon, and Mihaela van der Schaar. DeepHit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 2314–2321, 2018. doi: 10.1609/aaai.v32i1.11842.

Giwoong Lee, Eunho Yang, and Sung Hwang. Asymmetric multi-task learning based on task relatedness and loss. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pp. 230–238, 2016.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 7294–7305. PMLR, 2020.

James Pickands, III. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131, 1975.

David Rindt, Robert Hu, David Steinsaltz, and Dino Sejdinovic. Survival regression with proper scoring rules and monotonic neural networks. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 1190–1205, 2022.

Jamshid Sourati, Murat Akcakaya, Todd K. Leen, Deniz Erdogmus, and Jennifer G. Dy. Asymptotic analysis of objectives based on Fisher information in active learning. *Journal of Machine Learning Research*, 18(34):1–41, 2017.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 9120–9132, 2020.

UCI Machine Learning Repository. Gas turbine CO and NOx emission data set. https://doi.org/10.24432/C5WC95, 2019. Accessed 2026-09-25; CC BY 4.0.

U.S. Geological Survey. Earthquake catalog: FDSN event web service. https://earthquake.usgs.gov/fdsnws/event/1/, 2026. Accessed 2026-09-24.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Fan Yang, Hongyang R. Zhang, Sen Wu, Christopher Ré, and Weijie J. Su. Precise high-dimensional asymptotics for quantifying heterogeneous transfers. *Journal of Machine Learning Research*, 26 (113):1–88, 2025.

Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

A Proofs

A.1 Proof of Theorem 1

Let s_C be the score of the common experiment. Differentiability in quadratic mean gives $\mathbb{E}s_C = 0$ and $\mathbb{E}s_C s_C^\top = I_c$. On the identifiable tangent space $\mathrm{Range}(I_c)$, the limiting covariance of an efficient estimator is I_c^+. By assumption, the target is locally a smooth function of those identifiable coordinates. The delta method then gives variance $g_a^\top I_c^+ g_a$.

For the endpoint experiment, the score for scalar p is $(Y-p)/\{p(1-p)\}$ and its Fisher information is $1/\{p(1-p)\}$. The Bernoulli MLE therefore has variance $p(1-p)/n$; applying the derivative $d\log p/dp = 1/p$ gives $(1-p)/(np)$.

It remains to prove the range condition. If $g_a \notin \mathrm{Range}(I_c)$, symmetry of I_c implies that there is $v \in \mathrm{Null}(I_c)$ with $g_a^\top v \neq 0$. Since $v^\top I_c v = \mathbb{E}(v^\top s_C)^2 = 0$, we have $v^\top s_C = 0$ almost surely. If a regular asymptotically linear estimator of $\log p_a$ based on C had influence function $\phi(C)$, pathwise differentiation would require $g_a^\top v = \mathbb{E}\{\phi(C)v^\top s_C(C)\} = 0$, a contradiction. Therefore no regular asymptotically linear estimator exists. More generally, quadratic-mean differentiability makes observations from the common experiment at θ and $\theta + v/\sqrt{n}$ asymptotically indistinguishable, whereas their target log-probabilities differ by $g_a^\top v/\sqrt{n} + o(n^{-1/2})$. A regular root-$n$ limit cannot accommodate this nonzero shift under indistinguishable experiments. □

Information under label refinement. Let F denote the full categorical label and $T(F)$ either coarsening. Under the usual interchange of differentiation and integration, the coarsened score is $\mathbb{E}\{s_F(F;\theta) \mid T(F)\}$. The conditional-variance decomposition gives $I_F - I_T = \mathbb{E}\{\mathrm{Var}(s_F \mid T)\} \succeq 0$. Equivalently, every estimator based on a coarsening is also available in the full experiment, so the efficient variance bound for an identifiable target cannot be larger under F. This establishes the scope statement in Section 2.

A.2 Proof of the rare-tail corollary

Eq. 1 and $1 - p_a \to 1$ give $\mathcal{A}_a \sim \{p_a g_a^\top I_c^+ g_a\}^{-1} \to \infty$. □

A.3 Proof of Proposition 1

Let $I = \sum_j \mu_j u_j u_j^\top$ be an orthonormal eigendecomposition. Expanding g in this basis gives the stated identity. Each zero eigenvalue contributes $(u_j^\top g)^2/\lambda$. Hence a nonzero null component forces divergence. If $g \in \mathrm{Range}(I)$, all zero-eigenvalue coefficients vanish and

$$g^\top I^+ g - V_\lambda(g) = \sum_{\mu_j > 0} \frac{\lambda (u_j^\top g)^2}{\mu_j(\mu_j + \lambda)} \leq \frac{\lambda}{\mu_+^2}\|g\|^2. \qquad \square$$

A.4 Proof of Proposition 2

Let $A = I + \lambda I_d$ and $\widehat{A} = \widehat{I} + \lambda I_d$. Both inverses have operator norm at most $1/\lambda$. Adding and subtracting $g^\top \widehat{A}^{-1} g$ gives

$$|\widehat{g}^\top \widehat{A}^{-1}\widehat{g} - g^\top \widehat{A}^{-1} g| \le (2G\epsilon_g + \epsilon_g^2)/\lambda.$$

The resolvent identity $\widehat{A}^{-1} - A^{-1} = \widehat{A}^{-1}(I - \widehat{I})A^{-1}$ then bounds the remaining quadratic form by $G^2\epsilon_I/\lambda^2$, proving Eq. 2. If $I \succ 0$, its eigendecomposition gives

$$0 \le V_0 - V_\lambda = \sum_j \frac{\lambda (u_j^\top g)^2}{\mu_j(\mu_j + \lambda)} \le \lambda G^2/\mu^2.$$

The triangle inequality proves the final claim. □

A.5 Proof of Proposition 3

Let $z_j = c_j - t > 0$ and $z_a = a - t$. Up to sign, the gradients of the log survival probabilities at c_1, c_2, and a with respect to $(r_\mathrm{p}, r_\mathrm{q}, c)$ are $(c_1, 0, 0)$, $(t, z_2, z_2^2/2)$, and $(t, z_a, z_a^2/2)$. The last vector is in the span of the first two only if $z_a = z_2$, which is excluded by $a > c_2$. The two thresholds therefore leave a component of the target gradient unidentified. Adding c_3 supplies $(t, z_3, z_3^2/2)$; the determinant of these three threshold rows is $c_1 z_2 z_3 (z_3 - z_2)/2 \neq 0$. The common-survival Jacobian is therefore full rank, and an interior categorical reparameterization preserves that rank. □

B Full protocol and evidence provenance

Frozen inputs. The manifest, original aggregate, replay adjudication, and lossless cell file have SHA-256 digests, respectively, `5b03ca64...fd8c0`, `8be069c1...bd05`, `2206ddff...d43d`, and `312cd65e...36a`. The reanalysis verifies these hashes before reading the 3,200 replay rows: 2 contexts, 4 probabilities, 2 sample sizes, and 200 seeds. Every count vector sums to N, and common-only estimates are invariant to reallocating counts within the unobserved target split. The immutable manifest calls the common estimator a categorical-count MLE, but the implementation uses the known one-half upper mass: $S(c_1) = 1 - 2N_1/n$ and $S(c_2) = 2(N_3 + N_4)/n$. The conditional MLE instead normalizes by $n - N_0$. We retain the implementation and use the correct term, moment estimator. Its variance, derived below, can exceed the efficient bound, so R^2 need not converge to $\mathcal{A}$.

Randomness. The original study uses NumPy PCG64DXSM with `SeedSequence([20260819, seed, h, p, N])` and seed IDs 2000–2199. The independent replication uses root 20260924 and IDs 5000–5199. The finite-budget study uses 500 seeds per cell; each curvature study uses 200. The original stress-test grid was specified in `protocol.md`. The follow-up information-completion design was specified separately in `revision_protocol.md` before random-number generation, with $c_3 = 4.0$ and `SeedSequence([20260924, 9000+seed, h, p, k])`. Seeds, cells, and metrics were retained as specified after observing the results.

Learned-model audit. The protocol `learned_audit_protocol.md` was hashed before running seeds 7000–7029. Three separate development seeds checked gradients and runtime. The protocol specifies the data-generating process, network width, optimizer, early stopping, two targets, 2,001-point evaluation grid, integrated absolute log error, and 20,000-resample paired bootstrap. All 30 seed rows and training diagnostics are saved; the primary target and metric are unchanged.

Prospective learned validation. Following development runs with seeds 999981–999985, the target grid, model-size pairs, indifference band, 24 study seeds, proper-score diagnostics, CET tuning grid, and success criterion were fixed in `prospective_audit_protocol.md`. Study seed IDs 8000–8023 use a disjoint root namespace. The protocol and three implementation files are hash-verified before rerunning. Every lossless training row, selected CET strength, and candidate validation loss is saved; all targets, model sizes, and seeds were retained after execution.

Estimated-Fisher pilot. The protocol `estimated_fisher_protocol.md` specifies the 20,000-observation common-label pilot, four-parameter likelihood, 1,000 nonparametric refits, simultaneous intervals, and decision rule. Pilot seeds are disjoint from all training streams. As reported in Section 4.4, three harm decisions and two abstentions fail the minimum-count and help/harm requirements. The JSON results contain pilot counts, fitted coefficients, refit diagnostics, intervals, and decisions for each model size.

Table 4: Common-label pilot audit. Intervals are nominal Bonferroni simultaneous 95% bootstrap percentile intervals for Q_a. The rule uses the practical band $[0.95, 1.05]$.

a	Plug-in Q_a	Nominal simultaneous 95% interval	Decision
4	0.48	[0.41, 0.56]	harm
5	0.49	[0.39, 0.64]	harm
6	0.58	[0.42, 0.83]	harm
7	0.74	[0.48, 1.17]	abstain
8	0.99	[0.59, 1.76]	abstain

Second external dataset. The UCI archive hash, recommended temporal split, thresholds, estimators, and the descriptive status of the follow-up analysis are documented in `external_validation_protocol.md`. The CC BY 4.0 source archive is checked against SHA-256 `55fdb1ac...29195` before reading any row.

Final sensitivity audit. The pre-specified `final_audit_protocol.md` assigns 20,000 paired nonparametric resamples and a PCG64DXSM stream to each of the twelve high-N primary cells. The fifth percentile of the resampled mean-error ratio gives the one-sided 95% sensitivity bound. All primary delta-method decisions are unchanged.

Inference. For m paired seed-level errors (E_i, S_i), $L = \log(\bar{E}/\bar{S})$ and

$$\widehat{\mathrm{Var}}(L) = \frac{1}{m}\left(\frac{s_E^2}{\bar{E}^2} + \frac{s_S^2}{\bar{S}^2} - \frac{2s_{ES}}{\bar{E}\bar{S}}\right).$$

The lower bound is $\exp\{L - 1.64485\,\mathrm{SE}(L)\}$. Slope variance is the weighted sum of pointwise log-ratio variances because the experiments at different probabilities use distinct random streams. The primary criterion is an intersection-union test: all components must pass, so no multiplicity adjustment is needed for the conjunction.

Table 5: Delta-method and paired-bootstrap one-sided 95% lower bounds. Every cell-level decision is unchanged.

h	p	Original delta	Original boot.	Rep. delta	Rep. boot.
-1	1.09e-04	1.01	1.01	0.88	0.88
-1	2.17e-05	1.71	1.71	1.66	1.67
-1	2.17e-06	4.25	4.26	3.65	3.65
1	1.09e-04	1.02	1.02	1.04	1.04
1	2.17e-05	1.75	1.75	1.70	1.70
1	2.17e-06	4.35	4.35	4.44	4.44

Table 6: Independent rarity-slope confirmation.

h	Original slope	Original LCB	Independent slope	Independent LCB
-1	0.845	0.742	0.841	0.734
$+1$	0.854	0.750	0.858	0.759

B.1 VARIANCE OF THE MOMENT ESTIMATOR

Write $s_j = S(c_j)$, $z = a - t$, $d = c_2 - t$, $b_1 = t(1 - z/d)/c_1$, and $b_2 = z/d$. Then $\log p_a = \log(1/2) + b_1 \log s_1 + b_2 \log s_2$. The implemented estimator substitutes $\widehat{s}_1 = 1 - 2N_1/n$ and $\widehat{s}_2 = 2(N_3 + N_4)/n$. For the two proportions $u = (N_1/n, (N_3 + N_4)/n)$, let $q = ((1 - s_1)/2, s_2/2)$ and $d_u = (-2b_1/s_1, 2b_2/s_2)^\top$. The multinomial delta method gives

$$V_{\mathrm{mom}} = d_u^\top \{\mathrm{diag}(q) - qq^\top\} d_u.$$

Consequently, its asymptotic endpoint-to-common variance ratio is $\mathcal{A}_{\mathrm{mom}} = (1 - p_a)/(p_a V_{\mathrm{mom}})$, rather than the efficient $\mathcal{A}_a$ of Theorem 1. Under the theorem's model, $V_{\mathrm{mom}} \geq g_a^\top I_c^+ g_a$. This distinction explains why the empirical comparison can differ from the efficient information bound even at large sample sizes.

C ADDITIONAL FINITE-BUDGET RESULTS

The Jeffreys estimator is $(K + 1/2)/(N + 1)$. Its positive estimate keeps log error finite when the endpoint count is zero. At p_3, the ratios of mean absolute log error for the Jeffreys endpoint estimate to those for the common-only estimate across expected counts $10, 20, 50, 100, 500, 2000$ are approximately 4.97, 5.11, 4.83, 5.02, 4.62, and 4.37 for $h = -1$, and 5.21, 4.82, 4.96, 5.48, 4.43, and 5.42 for $h = 1$. At p_1 ratios cluster near one, consistent with the unstable boundary cell. The JSON results contain all 36 rows, including zero-count frequencies and confidence bounds.

D INITIAL LEARNED-MODEL DIAGNOSTIC

The initial 30-seed experiment used width-16 endpoint-only, common-only, and full-label networks under the same optimizer, maximum-epoch budget, and early-stopping rule, but did not match their output-head parameter counts. All methods were scored on a fixed 2,001-point context grid.

Table 7: Initial learned-model stress test. Entries report mean integrated absolute log-probability error (SE); lower is better. Δ_c and Δ_f are paired endpoint-minus-common/full differences, with one-sided 95% bootstrap LCBs in brackets.

a	Endpoint	Common	Full	Δ_c [LCB]	Δ_f [LCB]
6	0.242 (0.018)	0.295 (0.032)	0.203 (0.019)	-0.054 [-0.115]	+0.039 [+0.009]
7	0.395 (0.030)	0.373 (0.040)	0.298 (0.030)	+0.022 [-0.052]	+0.097 [+0.042]

At $a = 7$, endpoint-only and common-only errors are 0.395 and 0.373, but the paired mean difference is 0.022 with lower bound -0.052, so the common-only comparison fails its primary criterion. At $a = 6$, common-only performs worse (0.295 versus 0.242). Full-label errors are 0.298 and 0.203; endpoint-minus-full lower bounds are 0.042 and 0.0087. Common-only outperforms endpoint-only on 16/30 and 13/30 seeds; full-label does so on 21/30 and 19/30. These results motivated the separately specified prospective study with approximately matched parameter counts in Section 4.3; the two studies are analyzed separately.

E EXTERNAL-DATA DETAILS

The USGS FDSN event-service query requested CSV output with `starttime=2000-01-01`, `endtime=2026-01-01`, latitude 32–42, longitude -125–-114, `minmagnitude=3.0`, and `eventtype=earthquake`. The saved raw file hash is `583b51bd...392d0`. We retain 11,991 reviewed events from 12,159 rows; the test years 2010–2025 contribute 8,219 held-out events. The catalog combines magnitude types; 9,337 raw rows are local magnitude and 1,898 are moment magnitude, with smaller groups of other types. The analysis retains these original magnitude types.

The common-only exponential estimator observes three bins, $[3, 3.5)$, $[3.5, 4)$, and $[4, \infty)$, and maximizes their multinomial likelihood. The target splits at 4.5, 5.0, and 5.5 are unobserved. The endpoint-only estimator uses the training target count. The exact-magnitude benchmark estimates

$\lambda = 1/\overline{(M-3)}$. All methods output one conditional exceedance probability per fold; their held-out score is Bernoulli log loss. Overlap among ten-year training windows motivates the descriptive interpretation.

Table 8: USGS event-weighted log-loss difference from the target-specific endpoint. Negative values favor the alternative estimator.

Target	Method	Δ log loss	Better folds	Total folds
4.5	Common bins	-1.3958e-05	6	16
4.5	Exact magnitude	-4.5305e-04	7	16
5.0	Common bins	+2.7308e-04	6	16
5.0	Exact magnitude	-6.7801e-05	7	16
5.5	Common bins	-1.5565e-04	7	16
5.5	Exact magnitude	-1.6451e-04	11	16

For UCI, the source files are already partitioned by year and sorted chronologically. The common categorical counts on 2011–2013 are $(13{,}565, 5{,}695, 2{,}041, 890)$ for bins split at 70, 80, and 90 mg/m^3. The fitted above-70 exponential rate is 0.11068. The endpoint and common-only training probabilities are respectively 0.01967 and 0.01405 at target 100, and 0.00782 and 0.00464 at target 110. All 7,158 rows from 2014 and 7,384 rows from 2015 are retained. The JSON results also include the exact-value comparator and Brier scores; Table 3 focuses on endpoint-only and common-only log loss.

F Exact marked-record oracle

For a discrete-time trajectory $X_{1:H}$ and boundaries $a, b > 0$, define $U_t = \max_{s \le t} X_s$ and $D_t = \max_{s \le t}(-X_s)$. A competing first-passage query returns upper, lower, or neither according to which boundary is first crossed by a requested horizon.

Proposition 4 (Record sufficiency and ancillary rendering)**.** *Every finite-horizon competing first-passage label is a deterministic function of $R_{1:H} = \{(U_t, D_t)\}_{t=1}^{H}$. If a renderer samples a trajectory Z from a parameter-independent kernel $K(\cdot \mid R)$ and satisfies $R(Z) = R$ almost surely, then all first-passage labels are preserved. Moreover the rendered and record experiments have identical Fisher information about θ.*

F.1 Proof of Proposition 4

The first upper crossing time is $\inf\{t \le h : U_t \ge a\}$ and the first lower crossing time is $\inf\{t \le h : D_t \ge b\}$, so the label is measurable with respect to R. If $R(Z) = R$, these two crossing times and their ordering are unchanged.

To compare Fisher information, factor the joint density as $f_\theta(r)K(z \mid r)$, with $r = R(z)$ recoverable from z and K independent of θ. Its score is therefore $\nabla_\theta \log f_\theta(R(z))$, exactly the record score. Taking the second moment proves equality of Fisher information. □

Unlike importance-sampling queries of fitted event-sequence models (Boyd et al., 2023) or monotone neural first-hitting-time prediction (Chen et al., 2025), this proposition establishes that the record representation preserves query answers and supports exact verification.

The state $(u, d) \in \{0, \dots, 8\}^2$ stores nondecreasing upper and lower records. A softmax selects no record, an upper jump, or a lower jump; conditional jump sizes have normalized exponentially decaying weights. For query (a, b, h), dynamic programming starts at $(0, 0)$ and transfers probability to the upper or lower outcome at the first boundary crossing. Probability remaining after h steps gives the neither outcome. The three probabilities sum to one within 10^{-12}.

All fourteen verification tests pass. They cover a handwritten trajectory/query truth table, exhaustive tiny-state path enumeration versus dynamic programming to 10^{-12}, randomized record/trajectory label agreement under four renderer modes, and normalization, calibration, and query semantics.