Residual Expertise Is Not Decision Value

Nidhish Shah *
Independent Researcher
shah@nidhish.dev

Shaurjya Mandal *
Harvard University
smandal2@mgh.harvard.edu

Asfandyar Azhar
University of Oxford
asfandyar.azhar@st-annes.ox.ac.uk

Abstract

Audits of human–AI complementarity target the wrong estimand. They ask whether the human knows more than the model; deployment asks whether that knowledge would change the action. Residual predictive expertise is not residual decision value. Deployment rewards actions, and a posterior movement that does not cross a reward-induced action boundary changes nothing; a signal can be informative under log-loss while leaving the model-only action optimal. We formalize the missing quantity as *boundary regret*: the regret of retaining the model-only action after observing the human signal. In any finite-action Bayesian decision problem, the value of consulting the human equals expected boundary regret, formally verified in Lean 4. The identity turns when humans help into a geometric question: which decision facets does the human signal move belief across? Complementarity is therefore reward-relative. A given human signal can be valuable under one reward matrix and worthless under another, the human's knowledge unchanged; the boundary moved, not the expertise. The gap is widest precisely where asymmetric costs dominate: high-stakes clinical and operational decisions. Audits built on residual predictive expertise overstate the value of human review and route scarce review toward cases where the human is informative but the action will not move. Boundary regret is the quantity to measure.[1]

1 Introduction

Human review is often justified by asking whether the human knows something the model does not. In clinical AI, a radiologist may catch findings that an imaging model misses, and residual-expertise audits can test whether the human signal carries predictive information beyond the model [Alur et al., 2023, 2024]. But deployment rewards actions, not predictive information. A radiologist can shift the posterior without changing the admit/discharge decision; only posterior movements that cross the reward-induced action boundary matter in practice. Residual predictive expertise and residual decision value are not the same estimand.

Complementarity is therefore reward-relative. The same human signal can matter under one reward matrix and not under another, not because the human's expertise changed, but because the boundary moved. Human–AI complementarity is not a property of the prediction problem alone. For deployment, it also depends on the model belief, the human signal, the action set, and the reward matrix. When audits treat residual predictive expertise as residual decision value, they can spend scarce review budget on cases where the human is informative but operationally irrelevant.

*Equal contribution.

[1] Code and Lean 4 proofs: github.com/nidhishs/residual-expertise-is-not-decision-value.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

We formalize the missing estimand as *boundary regret*: the regret of retaining the model-only action after observing the human signal. For model belief b_x, model-only action $a_x = \arg\max_a r_a \cdot b_x$ (optimal under b_x), human-updated belief $b_{x,h} = P(Y \mid x, h)$, and terminal value $V(b) = \max_a r_a \cdot b$,

$$\mathrm{BR}^*(x, h) \;=\; V(b_{x,h}) \;-\; r_{a_x} \cdot b_{x,h}, \tag{1}$$

which is positive exactly when a rival action becomes preferred under $b_{x,h}$. In any finite-action Bayesian decision problem, the value of observing the human signal equals expected boundary regret (Theorem 2), a form of the classical value-of-information identity [Radner and Stiglitz, 1984, de Lara and Gossner, 2020]:

$$\mathrm{VoI}(H \mid b_x) \;=\; \mathbb{E}_{h|x}\big[\mathrm{BR}^*(x, h)\big]. \tag{2}$$

Our contribution is to operationalize this identity as a model-conditioned, reward-local human–AI audit, yielding a per-case, per-reader score for allocating scarce review. Predictive gain without a reachable boundary crossing has zero decision value for the deployed action set. We formalize in Lean 4 with Mathlib [de Moura and Ullrich, 2021, The mathlib Community, 2020] both the zero-regret boundary condition and the value-of-information identity above. The paper develops this criterion in four steps:

- We define boundary regret and prove, in Lean 4, that $\mathrm{VoI}(H \mid b_x) = \mathbb{E}_{h|x}[\mathrm{BR}^*(x, h)]$, with a geometric facet-activation criterion that characterizes when boundary regret is strictly positive (Theorems 2–3).
- We validate the plug-in estimator BR on synthetic finite-action problems, including a strong-null condition where humans carry substantial predictive signal yet boundary regret is zero everywhere: predictive gain is not sufficient for decision value.
- We show on CheXpert that residual predictive expertise and residual decision value separate as a function of reward geometry, exactly as the theory predicts: the same model, the same readers, and the same conditions yield opposite-signed correlations under different rewards.
- We show that ex ante BR improves scarce-review allocation over reward-aware margin, reward-agnostic uncertainty, and residual-expertise baselines where actionable mass is dense, and converges to reward-aware margin where it is sparse or absent (Theorem 3).

2 Boundary Regret

2.1 Decision Setup

We work in finite-action Bayesian decision problems [DeGroot, 1970, Howard, 1966]. An instance with features $x \in \mathcal{X}$ has unknown state $y \in \mathcal{Y} = \{1, \ldots, K\}$. The decision-maker chooses an action a from a finite set $\mathcal{A}$, $|\mathcal{A}| \geq 2$, and receives reward $r_a(y)$ specified by $R \in \mathbb{R}^{|\mathcal{A}| \times K}$. A *belief* $b \in \Delta^{K-1}$ is a probability vector over states; write $b_x = P(Y \mid x)$ for the model belief at instance x. At any belief, the terminal value is the largest expected reward available from the action set. The model-only action a_x attains this value under b_x. Each action has a decision region: the beliefs at which that action is optimal. Formally,

$$V(b) \;=\; \max_{a \in \mathcal{A}} r_a \cdot b, \qquad a_x \in \arg\max_{a \in \mathcal{A}} r_a \cdot b_x, \qquad \mathcal{R}_a \;=\; \{b : r_a \cdot b \geq r_{a'} \cdot b \;\forall a' \in \mathcal{A}\}. \tag{3}$$

Each $\mathcal{R}_a$ is a convex polytope, and the relative boundary $\mathcal{R}_a \cap \mathcal{R}_{a'}$ is a *decision facet*. The regret of action a at belief b is $\mathrm{regret}(a, b) = V(b) - r_a \cdot b \geq 0$, with equality iff a is optimal.

2.2 Boundary-Regret Estimand

After observing the human signal h, the model belief b_x updates to $b_{x,h}$. That update has decision value only if it crosses a decision facet, making some rival action better than a_x.

Definition 1 (Boundary Regret)**.** *For an instance x with model-only action a_x and human signal h, the* instance-level boundary regret *is the regret of retaining a_x after the belief updates from b_x to $b_{x,h}$. The* ex ante boundary regret *is the expected decision value of the human signal at instance x, before h is observed. Formally,*

$$\mathrm{BR}^*(x, h) \;=\; V(b_{x,h}) \;-\; r_{a_x} \cdot b_{x,h}, \qquad \mathrm{BR}^*(x) \;=\; \mathbb{E}_{h|x}[\mathrm{BR}^*(x, h)]. \tag{4}$$

We call $\mathrm{BR}^*(x, h)$ the oracle *boundary regret;* BR *denotes its plug-in estimate (§2.5).*

2.3 Decision-Value Identity

The value of observing the human signal is exactly expected boundary regret, a form of the classical value-of-information identity [Radner and Stiglitz, 1984, de Lara and Gossner, 2020]. This rewrites value of information as the loss from keeping the model-only action after the posterior update.

Theorem 2 (VoI as Expected Boundary Regret)**.** *In any finite-action Bayesian decision problem with a_x optimal at b_x, where $b_{x,h}$ is the Bayesian posterior update of b_x after observing h, the value of observing the human signal equals expected boundary regret:*

$$\mathrm{VoI}(H \mid b_x) \;:=\; \mathbb{E}_{h|x}\big[V(b_{x,h})\big] - V(b_x) \;=\; \mathbb{E}_{h|x}\big[\mathrm{BR}^*(x,h)\big] \;=\; \mathrm{BR}^*(x). \tag{5}$$

Proof. Use the Bayesian martingale identity: the posterior averages back to the prior. Therefore the expected reward of any fixed action is unchanged after averaging over the posterior update:

$$\mathbb{E}_{h|x}[b_{x,h}] = b_x, \qquad \mathbb{E}_{h|x}[r_a \cdot b_{x,h}] = r_a \cdot b_x \quad \text{for any fixed action } a.$$

The left-hand side is the definition of value of information. Then use the identity above and the optimality of a_x at b_x. Subtract the fixed-action value from the posterior value:

$$\mathbb{E}_{h|x}[V(b_{x,h})] - V(b_x) = \mathbb{E}_{h|x}[V(b_{x,h})] - \mathbb{E}_{h|x}[r_{a_x} \cdot b_{x,h}] = \mathbb{E}_{h|x}[V(b_{x,h}) - r_{a_x} \cdot b_{x,h}]. \quad \square$$

2.4 Boundary Crossing

Boundary regret is positive exactly when at least one reachable human-updated posterior leaves the incumbent decision region. If every reachable posterior stays in $\mathcal{R}_{a_x}$, the human signal may improve prediction, but the model-only action remains optimal.

Theorem 3 (Facet Activation Bridge)**.** *For a coherent Bayesian belief/update pair $(b_x, b_{x,h})$, there is positive ex ante boundary regret if and only if some reachable posterior strictly prefers a rival action to the model-only action. Equivalently, boundary regret is zero if and only if every reachable posterior remains in the model-only action's decision region:*

$$\begin{aligned} \mathrm{BR}^*(x) > 0 \quad &\iff \quad \exists h,\ a' \neq a_x \text{ with } P(h \mid x) > 0 \text{ and } r_{a'} \cdot b_{x,h} > r_{a_x} \cdot b_{x,h}, \\ \mathrm{BR}^*(x) = 0 \quad &\iff \quad b_{x,h} \in \mathcal{R}_{a_x} \text{ for every } h \text{ with } P(h \mid x) > 0. \end{aligned} \tag{6}$$

Proof. By Theorem 2, ex ante boundary regret is the expectation of the non-negative instance-level regret. Thus $\mathrm{BR}^*(x) > 0$ iff some reachable h has positive instance-level boundary regret. By definition, this means $V(b_{x,h}) > r_{a_x} \cdot b_{x,h}$, which holds iff some action a' achieves $r_{a'} \cdot b_{x,h} > r_{a_x} \cdot b_{x,h}$. The converse follows by contraposition: if a_x is optimal at every reachable posterior, then $\mathrm{BR}^*(x,h) = 0$ for all h, hence $\mathrm{BR}^*(x) = 0$. $\square$

Corollary 4 (No Crossing, No Decision Value)**.** *If every reachable human-updated posterior $b_{x,h}$ lies in $\mathcal{R}_{a_x}$, then $\mathrm{VoI}(H \mid b_x) = 0$, regardless of how much H improves predictions as measured by log-loss, calibration, or any proper scoring rule.*

2.5 Deployment and Estimation

Potential vs. Realized Value. $\mathrm{BR}^*(x)$ is an information-theoretic ceiling on the per-instance decision value of review using the specified human signal. To realize that value, a system still needs correct elicitation, calibrated posterior updating, and operator compliance [Vaccaro et al., 2024, Buçinca et al., 2021].

Estimation. The identities above are exact for ideal Bayesian beliefs: b_x and $b_{x,h}$ are a coherent prior/posterior update pair. Empirically, we replace these objects with estimates. The estimated model belief determines a_x; a cross-fitted human-augmented model estimates the updated belief $\hat{b}_{x,h} \approx P(Y \mid x, h)$.

Thus BR is a plug-in estimate of boundary regret. For ex ante review allocation, we also estimate $P(h \mid x)$ before human review and compute

$$\widehat{\mathbb{E}}_{h|x}\Big[V(\hat{b}_{x,h}) - r_{a_x} \cdot \hat{b}_{x,h}\Big]. \tag{7}$$

Calibration affects how accurately this plug-in score recovers true boundary regret; Appendix B.1 reports a targeted binary stress test of systematic belief error. Section 3 also details how this quantity is estimated for each experiment.

3 Experiments

We have three sequential experiments. Section 3.1 validates the boundary-regret estimator in synthetic settings with closed-form ground truth. Section 3.2 tests whether residual predictive expertise and residual decision value diverge on a real human–AI dataset across reward structures. Section 3.3 asks whether an ex ante boundary-regret score allocates scarce human review better than standard baselines under asymmetric rewards.

3.1 Experiment 1: Estimator and Geometry Validation

Question. Does the plug-in estimator recover known boundary regret, and remain zero when a safe action dominates the reachable belief region?

Setup. We simulate finite-state Bayesian decision problems with known prior, reward matrix, and channels $P(m \mid y), P(h \mid y)$ ($M \perp H \mid Y$). For each sampled (y, m, h) we compute the exact posteriors $b_m = P(y \mid m)$ and $b_{m,h} = P(y \mid m, h)$, the model-only action a_m, and instance-level boundary regret

$$\mathrm{BR}^*(m, h) \;=\; V(b_{m,h}) - r_{a_m} \cdot b_{m,h}.$$

The estimator does not see the true channels: it estimates $\hat{P}(m \mid y), \hat{P}(h \mid y), \hat{p}(y)$ from finite data via 5-fold cross-fitting with Laplace smoothing, then computes plug-in BR on held-out instances ($N = 10{,}000$ per condition).

Conditions. The grid varies $G \in \{2, 3, 5\}$ decision groups and $K \in \{G, 2G\}$ latent states, covering the minimal boundary case ($G = 2$), the smallest genuinely multi-action case ($G = 3$), and a moderate stress test ($G = 5$); $K = 2G$ checks that estimation does not degrade when multiple latent states share a decision group. Three roles: C1 is a positive control (actions reward state groups, human signal crosses boundaries), C2 a weak null (prior deep inside one decision region, weak human channel), and C3 a strong null (highly predictive human signal but a safe action dominates every reachable posterior). For C3 we set $\tau = (1 + U)/2$, where U is the maximum reachable posterior mass on any decision group under the exact channels; $\tau > U$ guarantees no reachable posterior can make a non-safe action optimal while the human signal retains its predictive accuracy.

Results. Table 1 reports metrics averaged across the six configurations; rank correlation is reported only where BR^* varies (in both nulls $\mathrm{BR}^* \equiv 0$, so we report false-positive rate instead).

Table 1: Synthetic validation of BR averaged over $G \in \{2, 3, 5\}$, $K \in \{G, 2G\}$. C3's ΔLL (0.595) exceeds C1's because the human signal is strongly predictive of group membership, yet no posterior crosses the certified threshold. ΔLL: log-loss improvement from b_m to $b_{m,h}$.

Condition	Role	$\rho(\mathrm{BR}, \mathrm{BR}^*)$	FPR	$\%\,\mathrm{BR}^* > 0$	$\%\,\mathrm{BR} > 0$	ΔLL
C1	Positive control	0.901	0.000	37.9	37.9	0.408
C2	Weak null	n/a	0.000	0.0	0.0	0.019
C3	Strong null	n/a	0.000	0.0	0.0	**0.595**

The estimator separates predictive value from decision value. In C1, BR ranks true boundary regret well ($\rho = 0.901$) across the 37.9% of instances where crossings occur (per-configuration details in Appendix B); under the unshifted cross-fitted estimates, FPR is zero in both nulls. C3 is the critical case: the human signal achieves the highest log-loss improvement of any condition ($\Delta\mathrm{LL} = 0.595$), yet BR assigns zero boundary regret to every instance because no posterior crosses the certified threshold. High predictive accuracy on task-relevant features does not imply decision value when a safe action dominates the reachable belief region.

A paired binary stress test shifts the log-odds of $\hat{b}_x$ or $\hat{b}_{x,h}$ by ± 1 while holding the simulated cases, fitting protocol, and all other quantities fixed (Appendix B.1). The resulting FPR is component- and geometry-dependent: shifts in $\hat{b}_x$ yield 49.1–50.9% FPR in C1, whereas shifts in $\hat{b}_{x,h}$ yield 23.0–40.4% FPR in the C3 strong null; the complementary shifts remain zero in these canonical binary cells.

3.2 Experiment 2: Residual Expertise Does Not Equal Decision Value

Question. Does the correlation between residual predictive expertise and boundary regret invert across reward regimes on the same model, readers, and data?

Setup. We pair a pretrained DenseNet-121 chest X-ray classifier [Huang et al., 2017, Cohen et al., 2022] with five board-certified radiologists of the CheXpert reader study [Irvin et al., 2019] across five official competition conditions (Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion), yielding $N = 25$ reader–task pairs on 500 frontal radiographs. For each pair, the ground-truth label y is the majority vote of the *remaining four* readers, which eliminates structural h–y contamination. Model beliefs b_x are temperature-scaled [Guo et al., 2017] per pair on a 30% calibration split; evaluation uses the remaining 70% ($\approx$350 images per pair). The human-augmented posterior $\hat{b}_{x,h} \approx P(Y = 1 \mid x, h)$ is estimated via 5-fold cross-fitted logistic regression on features $[\mathrm{logit}(b_x),\ h,\ h \cdot \mathrm{logit}(b_x)]$. Spearman correlations between residual predictive expertise (log-loss gain over b_x) and BR are reported across the 25 pairs under each reward regime; label protocol details in Appendix C.

Reward Regimes. We evaluate three reward matrices $r_{a,y}$, each placing the decision threshold $r_{0,0}/(r_{0,0} + r_{1,1})$ [Pauker and Kassirer, 1980] in a distinct region of the model belief distribution. R1 is the identity ($r_{0,0} = r_{1,1} = 1$, errors 0), threshold $1/2$, in the bulk. R2 ($r_{0,0} = 1$, $r_{1,1} = 5$, errors 0) penalises false negatives five times as heavily as false positives, threshold $1/6$, below the bulk. R3 mirrors R2 ($r_{0,0} = 5$, $r_{1,1} = 1$), threshold $5/6$, in the sparse upper tail. R3 is the critical negative control: the geometry in Theorem 3 motivates the prediction that when the threshold sits in the tail, facet crossings become rare, boundary-regret mass collapses to near zero, and residual predictive expertise decouples from decision value regardless of how much the reader improves the model's log-loss.

Table 2: Boundary-regret diagnostics on $N = 25$ reader–task pairs: actionable BR mass collapses from R1 to R3 and BR concentrates near the active facet in all regimes. n_{v}: pairs where $\mathrm{BR} \not\equiv 0$; $\bar{\rho}_{\mathrm{facet}}$: mean Spearman between facet distance and BR.

Reward	Role	$\bar{\rho}_{\mathrm{LL}}$	$\rho{>}0/n_{\mathrm{v}}$	$\bar{\rho}_{\mathrm{facet}}$	Top-decile zero-BR	n_{v}
R1 (thr. $1/2$)	Primary result	$+\mathbf{0.477}$	$25/25$	-0.599	11.2%	25
R2 (thr. $1/6$)	Intermediate	-0.051	$15/25$	-0.866	28.3%	25
R3 (thr. $5/6$)	Negative control	-0.112	$0/7$	-0.221	$\mathbf{97.4}\%$	7

Reward-Conditional Dissociation. The correlation between residual predictive expertise (log-loss gain over the model alone) and boundary regret inverts with the reward (Table 2): $\bar{\rho}_{\mathrm{LL}} = +0.477$ under R1 (25/25 pairs positive), -0.051 under R2 (15/25), and -0.112 under R3 (0/7; one-sided sign test $p = 0.008$, with the remaining 18 of 25 pairs having $\mathrm{BR} \equiv 0$). Same model, same readers, same data; only the reward changes.

The mechanism is geometric. Under R1, threshold $1/2$ sits in a dense region of model beliefs, so a single reader update routinely crosses a facet and predictive expertise aligns with decision value. Under R3, the false-positive facet sits in the tail: only 7 of 25 pairs carry any non-zero BR, and, on average across pairs, 97.4% of top-disagreement-decile cases (90th percentile of model–reader disagreement $1 - b_x[h]$) have $\mathrm{BR} = 0$. The R3 correlation $\bar{\rho} = -0.112$ on these 7 pairs is directionally consistent with the geometric prediction; the collapse of n_{v} from 25 to 7 shows that non-zero boundary regret is confined to fewer pairs. R2 is intermediate (28.3% top-decile zero-BR). Facet distance correlates negatively with BR under every reward ($-0.599, -0.866, -0.221$ for R1, R2, R3): boundary regret concentrates near the active facet. R2's stronger concentration follows from the same geometry: with the threshold at $1/6$ nearly all decisions default to treat, so facet crossings tend to come from beliefs already near $1/6$. R1's centrally located threshold produces weaker concentration. We call this *reward-conditional dissociation*: a consequence of reward geometry, with the fitted beliefs held fixed.

Appendix D shows the ordering of mean correlations holds across in- and off-domain pretrained checkpoints (varying the belief geometry) and readers not involved in ground-truth labelling. As a negative control, Appendix E evaluates CIFAR-10H: on highly accurate classifiers whose softmax

probabilities rarely straddle a reward facet, $\bar{\rho}$ collapses to near-zero across all rewards ($+0.015$ under R1 vs. $+0.477$ on CheXpert R1), consistent with the geometric account in Theorem 3.

Implication for Predictive Audits. A residual-expertise audit in the style of Alur et al. [2023, 2024] would find substantial predictive value here (mean log-loss gain 0.387 nats across all 25 pairs). Under R3, 97.4% of top-disagreement-decile cases have zero estimated boundary regret, on average across pairs. The *estimand shift* from predictive to decisional complementarity is not a refinement; it changes the criterion for allocating human review. Section 3.3 tests this directly.

3.3 Experiment 3: Review Allocation under Scarce Human Input

Question. Given a fixed human-review budget, can an ex ante boundary-regret score identify the cases where reader input will improve realized decision utility, rather than improve prediction or flag model uncertainty? The prediction is regime-conditional: it should beat reward-aware but human-blind boundary proximity when reader updates can cross reward-relevant decision boundaries, and converge to it when such crossings are sparse or absent.

Setup. We use the same 25 reader–task pairs as in Section 3.2. For each case, the model-only belief b_x induces a model-only action a_x. For each review budget $q \in \{5\%, 10\%, 20\%, 50\%\}$, a policy ranks cases using only pre-review information. Reviewed cases observe the reader response h and take the human-updated action $\arg\max_a r_a \cdot \hat{b}_{x,h}$; unreviewed cases retain a_x. Utility gain is the held-out realized reward improvement over this no-review baseline. All learned policy scores are 5-fold cross-fitted on the same fold structure used to estimate $\hat{b}_{x,h}$, so review priority is always computed out of fold. Appendix G isolates the routing consequences of reward misspecification.

Policies. The proposed policy, BR, is an ex ante boundary-regret score: it predicts possible reader responses from x, predicts the corresponding human-updated posterior, and scores a case by the expected decision value of that update under the current reward. We compare it against six baselines. **Margin** is a reward-aware distance to the active decision facet, our primary comparison baseline: it captures reward geometry without estimating the human signal $P(h \mid x)$. **Residual** is a residual-expertise score, the deployable analogue of a predictive audit [Alur et al., 2023]. **Entropy** is reward-blind model uncertainty. These policies span two axes: Margin is reward-aware but human-blind, Residual is human-aware but reward-blind, Entropy is blind to both, and BR is both reward-aware and human-aware. **L2D** is the learned full-coverage competitor: like BR it is both reward-aware and human-aware, but reaches that coverage end-to-end through a cross-fitted logistic deferral classifier with reward-sensitive labels $\mathbb{1}[R[a_{x,h}, y] > R[a_x, y]]$, in the spirit of [Mozannar and Sontag, 2020, Mao et al., 2023], rather than through the reward's geometry and the reader-channel posterior. **Random** is the uninformed baseline. **Oracle** is the hindsight upper bound using realized labels.

Regime-Conditional Allocation. We evaluate all policies on the same 25 reader–task pairs under each reward matrix. Table 3 tests whether the dissociation in Section 3.2 becomes actionable under a scarce review budget. It does, but only where reward geometry leaves actionable mass reachable: BR beats Margin at every R1 budget, leads or ties across most R2 budgets, and has no systematic advantage in the R3 negative control. The paired wins, losses, and ties show when BR improves on Margin.

Dense actionable mass (R1). BR wins decisively at every budget (sign-test $p < 0.001$); decisive-win rates over Margin are 91%, 100%, 96%, 100% (tallies $20/2/3$, $24/0/1$, $22/1/2$, $25/0/0$). With actionable mass dense in the belief distribution, the reader channel adds signal beyond reward-aware geometric proximity: Margin finds cases near the active facet; BR identifies which of those the reader can move profitably. BR also achieves higher mean utility than both tested L2D implementations at every budget, with a substantially smaller gap for the shallow-tree classifier (Appendix F.2).

Sparse actionable mass (R2). The low threshold places most cases on the positive action; sparse boundary-regret mass concentrates near the active facet (exactly where Margin sorts), making Margin a strong baseline. All deployable reward-aware policies obtain much larger absolute gains than under R3. BR leads or ties Margin at every budget up to $q = 20\%$ with at most two paired losses ($3/0/22$, $3/0/22$, $5/2/18$); the $q = 50\%$ row is essentially a tie ($7/9/9$, mean difference 0.005).

Table 3: Ex ante review allocation under scarce human input. Entries report mean held-out utility gain over no review on $N = 25$ reader–task pairs. BR beats reward-aware Margin where actionable mass is dense (R1), leads or ties where mass is sparse (R2), and has no systematic edge where mass is absent (R3, negative control). W/L/T gives paired wins/losses/ties of BR against Margin; bold marks the best deployable policy in each row (Oracle excluded).

Reward	q	BR	Margin	Entropy	Residual	L2D	Random	Oracle	W/L/T
R1	5%	**+0.0459**	+0.0395	+0.0395	+0.0209	+0.0152	+0.0184	+0.0514	20/2/3
R1	10%	**+0.0883**	+0.0743	+0.0743	+0.0455	+0.0289	+0.0392	+0.1000	24/0/1
R1	20%	**+0.1638**	+0.1438	+0.1438	+0.0915	+0.0697	+0.0761	+0.2000	22/1/2
R1	50%	**+0.3325**	+0.2797	+0.2797	+0.2544	+0.2485	+0.2022	+0.4495	25/0/0
R2	5%	**+0.0514**	+0.0490	+0.0429	+0.0330	+0.0477	+0.0313	+0.0514	3/0/22
R2	10%	**+0.0999**	+0.0965	+0.0768	+0.0610	+0.0946	+0.0534	+0.1000	3/0/22
R2	20%	**+0.1914**	+0.1889	+0.1486	+0.1286	+0.1879	+0.1089	+0.2000	5/2/18
R2	50%	+0.4113	**+0.4161**	+0.2955	+0.3069	+0.4134	+0.2664	+0.4865	7/9/9
R3	5%	**+0.0029**	+0.0027	+0.0000	+0.0000	−0.0008	+0.0006	+0.0072	1/4/20
R3	10%	+0.0025	**+0.0031**	+0.0000	+0.0000	+0.0008	+0.0007	+0.0072	0/4/21
R3	20%	+0.0025	**+0.0029**	+0.0000	+0.0000	+0.0017	+0.0006	+0.0072	1/4/20
R3	50%	+0.0027	**+0.0029**	+0.0000	+0.0000	+0.0023	+0.0021	+0.0072	1/2/22

Near-zero actionable mass (R3). The active facet sits in the sparse upper tail of model beliefs; eighteen of twenty-five pairs have $\mathrm{BR} \equiv 0$ across the evaluation set. Allocation gains are correspondingly tiny: BR and Margin differ by less than 0.0006 at every budget, with only three to five decisive pairs out of 25.

Human Channel versus Geometric Proximity. Reward-aware distance to the active facet is a strong baseline: it already identifies cases near the decision boundary. The decisive-win rates over Margin under R1 show that BR's estimate of $P(h \mid x)$ and $P(Y \mid x, h)$ add reliable signal on top of that geometric distance where actionable mass is dense. Where boundary-regret mass is sparse (R2) or near zero (R3), BR and Margin yield similar utility gains. Residual and Entropy clarify why both axes matter: human information without reward geometry, or model uncertainty without either reward geometry or the reader channel, allocates review less efficiently. Residual uses the same cross-fitted response and posterior estimates as BR but ranks by predictive information rather than reward-relative regret. These comparisons are complementary but do not provide an additive decomposition of the utility gain.

4 Related Work

Four literatures bear most directly on boundary regret: residual human expertise, deferral and selective prediction, realized human–AI collaboration, and the decision-theoretic value of information.

Residual Human Expertise Audits. A line of work asks whether human judgments carry predictive information an algorithm has missed. Alur et al. [2023] test for residual expertise by checking whether expert predictions are independent of outcomes conditional on the available features; Alur et al. [2024] give an algorithmic-indistinguishability account of the same question. Kleinberg et al. [2018] formalize the selective-labels problem: when outcomes are observed only for cases selected by the human decision-maker, estimating how outcomes depend on both the instance and the human judgment requires accounting for non-random selection. Guo et al. [2026] push the question to the feature level via ILIV-SHAP, attributing each input feature's contribution to the agent-complementary VoI of an AI prediction. What these papers ask is whether humans carry residual predictive information, and which features drive it. We ask something narrower and decision-relevant: whether that information actually changes the reward-optimal action. Residual expertise can sit comfortably alongside zero residual decision value, and a feature-level decomposition does not tell you whether any of its predictive contributions cross a reward-induced decision facet.

Deferral, Selective Prediction, and Reject Options. Learning-to-defer methods train systems to choose between acting automatically and handing an instance to a human expert [Madras et al., 2018, Mozannar and Sontag, 2020, Hemmer et al., 2023, Mao et al., 2023]. Selective-prediction and reject-option methods address the related problem of when a classifier should abstain under a specified loss

[Ramaswamy et al., 2018, Franc et al., 2023], and cost-sensitive variants fold asymmetric penalties into the deferral objective [Alves et al., 2024]. All of these decide *when* to defer or abstain given a loss, an expert channel, and a collaboration protocol. The question we want to answer comes earlier: whether the human signal can make the model-only action suboptimal under the deployment reward, before any such protocol is on the table.

Realized Human–AI Collaboration. A separate body of work studies when teams actually achieve complementarity in deployment. Vaccaro et al. [2024] synthesize evidence that human–AI combinations often fail to outperform the better standalone agent. Peng et al. [2025] show that, in binary classification, calibration alone does not prevent a deterministic combination of forecasts from being less accurate than every individual agent. Schoeffer et al. [2024] show that explanations can change reliance without necessarily helping users identify correct advice, while Guo et al. [2024] propose a decision-theoretic framework that decomposes observed deployment loss into misreliance and signal-perception components. McLaughlin and Spiess [2024] model recommendation design for realized complementarity, and Hemmer et al. [2025] organize the empirical evidence on when complementarity appears. A related critical strand argues that complementarity also turns on action-set design, evaluation metrics, and interaction design [Straitouri et al., 2026]. That work is about teams once they are deployed: whether they realize value, what protocols are achievable in principle, how observed loss decomposes after the fact. Our question runs upstream of all of that. Is reward-local decision value in the joint distribution to begin with, before strategy choice, behavioral frictions, interfaces, and incentives have any chance to surface it or hide it?

Value of Information, Decision Geometry, and Decision-Focused Learning. Conceptually we are closest to value-of-information theory and decision geometry. Blackwell [1951] orders information structures by decision value. Radner and Stiglitz [1984] and de Lara and Gossner [2020] characterize when information value vanishes because posteriors stay inside a single action region, and Kamenica and Gentzkow [2011] expresses information value through convex value functions and induced decision regions. Decision-focused learning makes the adjacent algorithmic point that prediction should be trained for downstream decisions rather than generic accuracy [Elmachtoub and Grigas, 2022, Wilder et al., 2019], and decision calibration asks when predictions are indistinguishable to downstream best-response rules [Zhao et al., 2021]. In clinical statistics, decision curve analysis [Vickers and Elkin, 2006] evaluates predictive models under threshold-encoded reward rather than generic accuracy, plotting net benefit across a range of clinician decision thresholds; that work established the principle that model utility should be assessed under its downstream loss. Our contribution is to turn this geometry into a human–AI audit. For each case and reader, we estimate the expected reward gain from using the reader's information rather than retaining the model-only action. This gives a score for allocating scarce review by decision value rather than predictive expertise.

5 Conclusion

The audit question for human–AI complementarity is not whether the human carries information the model lacks but whether that information would change the reward-relevant action. We named the missing estimand boundary regret and proved that it equals the value of consulting the human in any finite-action Bayesian decision problem; that quantity is strictly positive precisely when some reachable posterior crosses a decision facet. The plug-in score this identity yields behaves as the bridge predicts: it allocates scarce review where decision-relevant mass is reachable and converges on reward-aware margin where it is not. Residual predictive expertise alone does not target this reward-relative value of review.

Limitations. Boundary regret characterizes the reward-local decision value latent in the joint distribution of (x, h, y). The results assume a fixed, finite action set and a reward matrix that does not change when the human signal is observed. Net workflow gains also depend on the elicitation interface, review completion, action adoption, and review cost [Vaccaro et al., 2024, Buçinca et al., 2021]. The deployed score is a plug-in estimate of an exact identity; systematic error in either the model belief or human-augmented posterior can change its magnitude and false-positive rate when it moves an estimate across a reward facet.

Broader impact. Routing scarce human review to cases that change the reward-relevant action raises operational efficiency in high-stakes settings such as clinical reading and triage. But the same machinery, fed a misspecified or contested reward, will under-allocate review for populations whose decision facets it does not reach. Any operational use of BR should pair the score with reward elicitation and a subgroup audit of which populations it removes human review from.

Outlook. Boundary regret does not single out the human as the information source. The same identity prices any candidate channel (a second-opinion reader, an additional diagnostic test, a more expensive feature) by what it would change about the reward-relevant action. A reward-faithful training signal for learning-to-defer follows: replace accuracy-based deferral labels with the regret of not consulting, and the construction extends to multi-expert and sequential-elicitation policies when the continuation-policy set is finite and fixed, with unchanged rewards and dynamics. A prospective deployment routing review on BR under operational rewards is the direct test of how much latent decision value a real workflow can extract. Complementarity is reward-relative. A signal can be informative under any proper scoring rule while leaving the model-only action optimal. The audit that ignores the boundary measures the wrong thing.

Acknowledgments and Disclosure of Funding

The authors received no third-party funding for this work and declare no competing interests.

References

Rohan Alur, Loren Laine, Darrick K. Li, Manish Raghavan, Devavrat Shah, and Dennis Shung. Auditing for human expertise. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 79439–79468, 2023. doi: 10.52202/075280-3477. arXiv:2306.01646.

Rohan Alur, Manish Raghavan, and Devavrat Shah. Human expertise in algorithmic prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 138088–138129, 2024. doi: 10.52202/079017-4384. arXiv:2402.00793.

Jean V. Alves, Diogo Leitão, Sérgio Jesus, Marco O. P. Sampaio, Javier Liébana, Pedro Saleiro, Mário A. T. Figueiredo, and Pedro Bizarro. Cost-sensitive learning to defer to multiple experts with workload constraints. *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=TAvGZm2Rqb`. arXiv:2403.06906.

David Blackwell. Comparison of experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 93–102. University of California Press, 1951.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021. doi: 10.1145/3449287.

Yaofo Chen. PyTorch CIFAR Models. GitHub repository, 2025. URL `https://github.com/chenyaofo/pytorch-cifar-models/tree/786c16252c0fc58ee9adac063f8337cc4a7a497a`. Revision 786c162 (2025-05-17).

Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P. Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. TorchXRayVision: A library of chest X-ray datasets and models. In *Proceedings of the 5th International Conference on Medical Imaging with Deep Learning*, volume 172 of *Proceedings of Machine Learning Research*, pages 231–249, 2022.

Michel de Lara and Olivier Gossner. Payoffs-beliefs duality and the value of information. *SIAM Journal on Optimization*, 30(1):464–489, 2020. doi: 10.1137/18M1230049.

Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*, volume 12699 of *Lecture Notes in Computer Science*, pages 625–635. Springer, 2021. doi: 10.1007/978-3-030-79876-5_37.

Morris H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.

Adam N. Elmachtoub and Paul Grigas. Smart "Predict, then Optimize". *Management Science*, 68 (1):9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Vojtech Franc, Daniel Prusa, and Vaclav Voracek. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023. arXiv:2101.12523.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330, 2017.

Ziyang Guo, Yifan Wu, Jason D. Hartline, and Jessica Hullman. A decision theoretic framework for measuring AI reliance. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 221–236, 2024. doi: 10.1145/3630106.3658901. arXiv:2401.15356.

Ziyang Guo, Yifan Wu, Jason Hartline, and Jessica Hullman. The value of information in human-AI decision-making. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2502.06152.

Patrick Hemmer, Lukas Thede, Michael Vössing, Johannes Jakubik, and Niklas Kühl. Learning to defer with limited expert predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6002–6011, 2023. doi: 10.1609/aaai.v37i5.25742. arXiv:2304.07306.

Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. Complementarity in human-AI collaboration: Concept, sources, and evidence. *European Journal of Information Systems*, 34(6):979–1002, 2025. doi: 10.1080/0960085X.2025.2475962.

Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966. doi: 10.1109/TSSC.1966.300074.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. URL `https://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html`. arXiv:1608.06993.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019. doi: 10.1609/aaai.v33i01.3301590.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101 (6):2590–2615, 2011. doi: 10.1257/aer.101.6.2590.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *Quarterly Journal of Economics*, 133(1):237–293, 2018. doi: 10.1093/qje/qjx032.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL `https://cave.cs.toronto.edu/kriz/learning-features-2009-TR.pdf`.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pages 6147–6157, 2018. arXiv:1711.06664.

Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 3578–3606, 2023. doi: 10.52202/075280-0159.

Bryce McLaughlin and Jann Spiess. Designing algorithmic recommendations to achieve human-AI complementarity. *arXiv preprint arXiv:2405.01484*, 2024.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 7076–7087, 2020. arXiv:2006.01862.

Stephen G. Pauker and Jerome P. Kassirer. The threshold approach to clinical decision making. *New England Journal of Medicine*, 302(20):1109–1117, 1980. doi: 10.1056/NEJM198005153022003.

Kenny Peng, Nikhil Garg, and Jon Kleinberg. A no free lunch theorem for human-AI collaboration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14369–14376, 2025. doi: 10.1609/aaai.v39i13.33574. arXiv:2411.15230.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9617–9626, 2019. URL `https://openaccess.thecvf.com/content_ICCV_2019/html/Peterson_Human_Uncertainty_Makes_Classification_More_Robust_ICCV_2019_paper.html`.

Roy Radner and Joseph E. Stiglitz. A nonconcavity in the value of information. In Marcel Boyer and Richard E. Kihlstrom, editors, *Bayesian Models in Economic Theory*, volume 5 of *Studies in Bayesian Econometrics*, pages 33–52. North-Holland, Amsterdam, 1984. ISBN 978-0-444-86502-1.

Harish G. Ramaswamy, Ambuj Tewari, and Shivani Agarwal. Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12(1):530–554, 2018. doi: 10.1214/17-EJS1388.

Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. Explanations, fairness, and appropriate reliance in human-AI decision-making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2024. doi: 10.1145/3613904.3642621.

Eleni Straitouri, Stratis Tsirtsis, Ander Artola Velasco, and Manuel Gomez-Rodriguez. Narrowing action choices with AI improves human sequential decisions. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 300 of *Proceedings of Machine Learning Research*, pages 1144–1152, 2026.

The mathlib Community. The Lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP)*, pages 367–381. ACM, 2020. doi: 10.1145/3372885.3373824.

Michelle Vaccaro, Abdullah Almaatouq, and Thomas W. Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12): 2293–2303, 2024. doi: 10.1038/s41562-024-02024-1.

Andrew J. Vickers and Elena B. Elkin. Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006. doi: 10.1177/0272989X06295361.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-Decisions pipeline: Decision-Focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1658–1665, 2019. doi: 10.1609/aaai.v33i01.33011658.

Shengjia Zhao, Michael P. Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. arXiv:2107.05719.

A Lean 4 Proof Artifacts

All formal results used in the paper are verified in Lean 4 using Mathlib. The development contains five modules: `Basics.lean`, `Regret.lean`, `JensenGap.lean`, `FacetActivation.lean`, and the umbrella import `ResidualExpertiseNotDecisionValue.lean`. Table 4 maps the paper's claims to Lean declarations.

Table 4: Map from paper statements to Lean 4 declarations.

Paper	Type	Lean declaration	Module
$V(b) = \max_a r_a \cdot b$	Def	`terminalValue`	Regret
$\mathrm{regret}(a, b) \geq 0$	Lem	`regret_nonneg`	Regret
$\mathrm{regret} = 0 \Leftrightarrow a$ optimal	Lem	`regret_eq_zero_iff`	Regret
$\mathbb{E}[b_{\mathrm{post}}] = b$	Lem	`prior_eq_weighted_posterior`	Basics
Jensen gap ≥ 0	Thm	`jensenGap_nonneg`	JensenGap
Thm 2 (VoI $= \mathbb{E}[\mathrm{BR}^*]$)	Thm	`jensenGap_eq_weighted_regret`	FacetActivation
Thm 3	Thm	`facet_activation_bridge`	FacetActivation

Primitives. `Basics.lean` defines `Belief`, `ObsKernel`, `marginalProb`, and `posterior`, and proves the posterior martingale identity `prior_eq_weighted_posterior`. `Regret.lean` defines `terminalValue` and `regret` and proves non-negativity and the zero-regret optimality characterization. `JensenGap.lean` defines Jensen gaps over posterior lotteries and proves non-negativity for belief-convex functions.

Paper Theorems. `FacetActivation.lean` contains the two formal claims used in the text. The weighted-regret identity `jensenGap_eq_weighted_regret` rewrites the value of information as expected regret of retaining the prior-optimal action after the posterior update. The bridge theorem `facet_activation_bridge` states that this expected regret is strictly positive exactly when some reachable posterior strictly prefers a rival action.

B Synthetic Validation: Per-Configuration Results

Table 5 reports per-configuration results across the three estimator-validation conditions of Section 3.1. False-positive rate is zero in every cell: across all 18 configurations the plug-in estimator never assigned positive boundary regret to an instance with $\mathrm{BR}^*(m, h) = 0$. Under the positive control C1, $\rho(\mathrm{BR}, \mathrm{BR}^*)$ degrades smoothly as the action/state space grows (G from 2 to 5, K from 2 to 10, giving $0.969 \to 0.745$), reflecting harder channel estimation, not a failure of the identity. Under the strong null C3, the human signal is by construction strongly predictive ($\Delta\mathrm{LL}$ up to 0.738), yet the estimator continues to assign zero boundary regret to every instance, the geometric prediction of Theorem 3.

B.1 Sensitivity to Systematic Belief Miscalibration

The zero FPRs above concern the unshifted cross-fitted estimates. To isolate the plug-in estimator's sensitivity to systematic belief error, we use the canonical binary cells ($G = K = 2$) of C1 and C3. C1 mixes positive- and zero-regret cases; C3 is the certified strong null with every case in the FPR denominator. This binary setting also matches the one-vs-rest downstream task without requiring an arbitrary multiclass direction of perturbation.

For a binary fitted belief $b = (b_0, b_1)$, we apply the signed log-odds offset

$$T_\delta(b)_1 = \frac{e^\delta b_1}{b_0 + e^\delta b_1}, \qquad T_\delta(b)_0 = \frac{b_0}{b_0 + e^\delta b_1}, \qquad \delta \in \{-1, +1\}.$$

Thus $\delta = +1$ multiplies the odds of state 1 by e, while $\delta = -1$ divides them by e. We shift $\hat{b}_x$ and $\hat{b}_{x,h}$ separately, retain the same simulated cases, fitting protocol, oracle boundary regret, and reward tolerance, and recompute only BR and FPR. The unshifted baseline is $\delta = 0$ and is already reported in the main experiment.

Table 5: Per-configuration results for synthetic estimator validation. G: decision groups; K: latent states; $|A|$: actions in the reward matrix. FPR is zero in every cell and is omitted; ρ is omitted when true boundary regret is constant zero (C2, C3).

Condition.	G	K	$\lvert A\rvert$	$\rho(\mathrm{BR}, \mathrm{BR}^*)$	$\%\,\mathrm{BR}^* > 0$	ΔLL
C1	2	2	2	0.969	35.5	0.226
C1	2	4	2	0.984	36.4	0.230
C1	3	3	3	0.941	37.9	0.403
C1	3	6	3	0.903	38.2	0.420
C1	5	5	5	0.862	39.8	0.577
C1	5	10	5	0.745	39.8	0.592
C2	2	2	2	n/a	0.0	0.006
C2	2	4	2	n/a	0.0	0.006
C2	3	3	3	n/a	0.0	0.015
C2	3	6	3	n/a	0.0	0.019
C2	5	5	5	n/a	0.0	0.035
C2	5	10	5	n/a	0.0	0.033
C3	2	2	3	n/a	0.0	0.447
C3	2	4	3	n/a	0.0	0.458
C3	3	3	4	n/a	0.0	0.592
C3	3	6	4	n/a	0.0	0.613
C3	5	5	6	n/a	0.0	0.723
C3	5	10	6	n/a	0.0	0.738

Table 6: Paired binary log-odds stress test of plug-in FPR. Each entry is the percentage of oracle-zero cases assigned $\mathrm{BR} > 0$ after shifting exactly one fitted belief. n_0 is the number of oracle-zero cases, using the same tolerance as FPR.

Condition	Baseline	$\hat{b}_x, -1$	$\hat{b}_x, +1$	$\hat{b}_{x,h}, -1$	$\hat{b}_{x,h}, +1$	n_0
C1	0.0	50.9	49.1	0.0	0.0	6,453
C3	0.0	0.0	0.0	40.4	23.0	10,000

The effect is component- and geometry-dependent. In C1, perturbing the model-only belief changes the incumbent action and creates false positives among its oracle-zero cases, whereas the same posterior-only shifts do not. In C3, the safe action makes the model-only shifts inert, but perturbing the human-augmented belief can falsely cross the safe-action facet. These one-at-a-time shifts intentionally need not preserve the Bayesian coherence relation between b_x and $b_{x,h}$: they are an input-level stress test of separately misspecified deployed estimates, not a perturbation of the exact identity.

This is one interpretable intercept-type stress magnitude, not a general robustness guarantee or an estimate of CheXpert calibration drift. Calibration fitted and evaluated on genuinely independent (or nested) data could reduce correctable systematic bias, but lower calibration error alone does not guarantee lower boundary-regret FPR near a reward facet. Standard conformal prediction instead targets coverage of prediction sets; it neither calibrates these point posteriors nor directly controls $\Pr(\mathrm{BR} > 0 \mid \mathrm{BR}^* = 0)$.

C CheXpert Protocol

Reader Study and Label Construction. The five board-certified radiologists (`bc1`, `bc2`, `bc3`, `bc5`, `bc7`) provided binary disease-presence ratings under the official CheXpert five-condition protocol (Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion) on the 500-image frontal-view test set. For each (condition, ground-truth reader) pair, the held-out label y is computed by *leave-one-out majority vote*: for reader r_i, $y = \mathbb{1}[\sum_{j \neq i} v_j \geq 3]$, where $v_j \in \{0, 1\}$ are the remaining four ground-truth readers' votes. This construction prevents structural leakage of h into y: the test reader is never a voter in their own ground truth. The three benchmark readers (`bc4`, `bc6`, `bc8`)

were not voters in the original study, so their pairs use all five ground-truth votes (equivalent to the official majority).

Model Belief and Calibration. Model probabilities b_x^{raw} are produced by the `torchxrayvision` DenseNet-121 [Huang et al., 2017, Cohen et al., 2022] with the standard X-ray preprocessing pipeline (8-bit normalisation, centre crop, 224×224 resize). The primary experiment uses the in-domain checkpoint `densenet121-res224-chex`; the robustness analyses below use all six publicly available checkpoints. For each (condition, reader) pair we apply per-pair logit-space temperature scaling: a stratified 30% calibration split (seed 42), with scalar T found by NLL minimisation in $[0.1, 10]$, gives the calibrated belief $b_x = \sigma(\text{logit}(b_x^{\text{raw}})/T)$.

Human-Augmented Posterior. On the remaining 70% evaluation split, we estimate $\hat{b}_{x,h} \approx P(Y = 1 \mid x, h)$ by 5-fold cross-fitted logistic regression on features $[\text{logit}(b_x),\ h,\ h \cdot \text{logit}(b_x)]$, with folds stratified on y (`StratifiedKFold`). The leave-one-out label construction prevents structural leakage of h into y; cross-fitting evaluates each case with a posterior model trained on the other folds, so its plug-in BR uses out-of-fold posterior predictions.

Evaluation. For each pair, all per-instance quantities (BR, log-loss gain, facet distance) are computed on the held-out fold of each instance and aggregated across folds. Pair-level summaries (Spearman ρ between log-loss gain and BR, top-decile zero-mass fraction) are then aggregated across the 25 pairs. Pairs with $\text{BR} \equiv 0$ across the entire evaluation set are excluded from the ρ computation; we report the count of valid pairs n_{v} alongside.

D Reward-Conditional Dissociation: Robustness

Table 7 extends the dissociation analysis of Section 3.2 to all six publicly available DenseNet checkpoints and all eight CheXpert radiologists ($6 \times 8 \times 5 = 240$ reader–task pairs). The ordering R1>R2≥R3 holds on every checkpoint individually: R1 is positive on every checkpoint ($+0.373$ to $+0.530$), R3 ranges from clearly negative (in-domain, -0.117 to -0.040) to essentially zero (off-domain NIH, $+0.007$), and R1 strictly dominates R2 throughout, with the gap compressing on off-domain pretraining. Off-domain pretraining (NIH, PadChest) attenuates the R1 signal and inflates R2, consistent with calibration drift moving belief mass toward the lower R2 facet at $1/6$. The direction of the dissociation is consistent with the geometric account.

Table 7: Reward-conditional dissociation across 6 DenseNet checkpoints $\times$ 8 readers $\times$ 5 conditions $= 240$ reader–task pairs. Entries report $\bar{\rho}_{\text{LL}}$, the mean per-pair Spearman correlation between log-loss gain and BR, under each reward (R1: thr. $1/2$; R2: thr. $1/6$; R3: thr. $5/6$). Bold marks the aggregate dissociation pattern across all 240 pairs.

Checkpoint	R1 $\bar{\rho}_{\text{LL}}$	R2 $\bar{\rho}_{\text{LL}}$	R3 $\bar{\rho}_{\text{LL}}$
In-domain pretraining			
`densenet121-res224-chex`	$+0.489$	-0.008	-0.117
`densenet121-res224-mimic_ch`	$+0.496$	$+0.197$	-0.040
`densenet121-res224-mimic_nb`	$+0.495$	$+0.210$	-0.042
`densenet121-res224-all`	$+0.530$	$+0.151$	-0.082
Off-domain pretraining			
`densenet121-res224-nih`	$+0.373$	$+0.313$	$+0.007$
`densenet121-res224-pc`	$+0.435$	$+0.313$	-0.050
All 240 pairs	$\mathbf{+0.470}$	$+0.196$	$\mathbf{-0.056}$

E CIFAR-10H: Negative Control

CIFAR-10H [Peterson et al., 2019] evaluates seven pretrained CIFAR-10 [Krizhevsky, 2009] classifiers [Chen, 2025] against 50 human annotators across 10 one-vs-rest binary tasks (3,500 reader–task pairs). Pretrained CIFAR-10 classifiers are highly accurate; their softmax probabilities rarely straddle a decision boundary for a given annotator's image subset. This motivates CIFAR-10H as a

negative control: we expect little actionable boundary-regret mass, consistent with the geometry in Theorem 3.

Table 8: CIFAR-10H negative control: mean Spearman $\bar{\rho}$ between log-loss gain and BR across 3,500 reader–task pairs.

Reward	$\bar{\rho}$ (valid)	$\bar{\rho}$ (all)	N valid	Top-decile BR $= 0$
R1 (thr. 0.500)	$+0.015$	$+0.007$	1,668	96.2%
R2 (thr. 0.167)	$+0.044$	$+0.024$	1,905	94.5%
R3 (thr. 0.833)	-0.008	-0.004	1,632	96.9%

The R1 correlation ($\bar{\rho} = +0.015$) is $\sim 32\times$ smaller than the primary CheXpert R1 correlation ($\bar{\rho} = +0.477$ across the 25 in-domain reader–task pairs; Section 3.2) and $\sim 30\times$ smaller than the 240-pair robustness average ($+0.470$; Table 7). In this negative control, even highly discordant reader responses usually leave the model-only action optimal under the fitted posteriors: mean zero-BR fractions in the top disagreement decile are 94.5–96.9% across rewards.

F Review Allocation: Robustness and Baseline Comparisons

F.1 Checkpoint and Reader Robustness

Table 9 extends the allocation experiment of Section 3.3 to all six DenseNet checkpoints and all eight CheXpert radiologists (240 reader–task pairs). Three patterns reproduce the main result. Under R1 (dense actionable mass), BR has higher mean utility than Margin at every budget ($p < 0.0001$ for $q \geq 10\%$). Under R2 (sparse but reachable mass), the pooled 240-pair comparison favors BR over Margin at $q \in \{5\%, 10\%, 20\%\}$ ($p < 0.0001$), compared with small, mostly tied differences on the 25 primary pairs. Under R3 (near-zero actionable mass), mean utility differences are small, with no systematic paired advantage in either direction (sign-test $p > 0.07$ at every budget; mean differences ≤ 0.0012 at $q = 5\%$ and ≤ 0.0004 for $q \geq 10\%$).

Table 9: Allocation robustness across 6 DenseNet checkpoints $\times$ 8 readers $\times$ 5 conditions $= 240$ reader–task pairs. Entries report mean held-out utility gain over no review. W/L/T gives paired wins/losses/ties of BR against Margin; bold marks the best deployable policy in each row (Oracle excluded).

Reward	q	BR	Margin	Entropy	Residual	L2D	Random	Oracle	W/L/T
R1	5%	**+0.0407**	+0.0403	+0.0403	+0.0192	+0.0101	+0.0195	+0.0513	106/74/60
R1	10%	**+0.0787**	+0.0763	+0.0763	+0.0413	+0.0247	+0.0393	+0.0996	134/71/35
R1	20%	**+0.1534**	+0.1444	+0.1444	+0.0939	+0.0647	+0.0788	+0.1931	167/51/22
R1	50%	**+0.3160**	+0.2936	+0.2936	+0.2653	+0.2267	+0.1947	+0.4132	198/25/17
R2	5%	**+0.0429**	+0.0350	+0.0368	+0.0186	+0.0258	+0.0215	+0.0604	112/25/103
R2	10%	**+0.0826**	+0.0665	+0.0685	+0.0385	+0.0541	+0.0426	+0.1078	121/23/96
R2	20%	**+0.1557**	+0.1252	+0.1276	+0.0870	+0.1123	+0.0849	+0.2013	132/22/86
R2	50%	**+0.3213**	+0.3052	+0.2682	+0.2430	+0.2826	+0.2079	+0.4399	110/98/32
R3	5%	**+0.0042**	+0.0030	+0.0001	+0.0008	+0.0025	+0.0002	+0.0115	22/31/187
R3	10%	**+0.0052**	+0.0048	+0.0001	+0.0013	+0.0043	+0.0003	+0.0120	18/30/192
R3	20%	+0.0055	**+0.0056**	+0.0001	+0.0020	+0.0049	+0.0005	+0.0120	13/23/204
R3	50%	+0.0057	**+0.0057**	+0.0003	+0.0025	+0.0050	+0.0017	+0.0120	8/18/214

F.2 Learning-to-Defer Classifier Choice

To assess whether the L2D comparison depends on classifier choice, we replace the logistic classifier with a shallow decision tree (maximum depth 3), holding the features, reward-sensitive training target, folds, and allocation protocol fixed. Both implementations use the same observations as BR: each fold trains on 280 cases and evaluates on 70 held-out cases. Table 10 uses the released implementation on the same folds as Table 3.

Under R1, L2D-Tree closes approximately 80–91% of the original mean-utility gap, although BR remains numerically higher at every budget. Under R2, the tree does not uniformly improve on logistic L2D; under R3, differences remain tiny and mixed. These comparisons concern the tested

Table 10: L2D implementation sensitivity on the 25 primary reader–task pairs. Entries are mean held-out utility gains over no review.

Reward	q	BR	L2D (logistic)	L2D-Tree
R1	5%	+0.0459	+0.0152	+0.0398
R1	10%	+0.0883	+0.0289	+0.0800
R1	20%	+0.1638	+0.0697	+0.1523
R1	50%	+0.3325	+0.2485	+0.3253
R2	5%	+0.0514	+0.0477	+0.0486
R2	10%	+0.0999	+0.0946	+0.0917
R2	20%	+0.1914	+0.1879	+0.1857
R2	50%	+0.4113	+0.4134	+0.3962
R3	5%	+0.0029	−0.0008	+0.0022
R3	10%	+0.0025	+0.0008	+0.0019
R3	20%	+0.0025	+0.0017	+0.0018
R3	50%	+0.0027	+0.0023	+0.0035

implementations at this sample size. A sufficiently expressive, consistent L2D learner targeting the same conditional expected utility can recover the same population ordering.

G Reward-Misspecification Sensitivity

Reward dependence is intrinsic to the decision problem. To isolate its routing consequences, we rank the same 25 CheXpert reader–task pairs by BR under R1 or R2, fitting the evidence models once per fold and holding them fixed across the two rankings. Both allocations use R2 actions, utility, and the same no-review baseline; only the cases selected for review differ.

Table 11: Routing-only reward sensitivity. Both columns report mean total utility under R2, not utility gain. Δ and W/L/T compare R1-ranked against R2-ranked allocation; overlap is the mean fraction of selected cases shared by the two rankings.

q	R1 ranking	R2 ranking	Δ	W/L/T	Overlap
5%	0.8073	0.8210	−0.0137	0/17/8	0.000
10%	0.8390	0.8688	−0.0298	1/20/4	0.000
20%	0.8942	0.9608	−0.0666	1/23/1	0.000
50%	0.9837	1.1807	−0.1970	0/25/0	0.345

R2's fivefold false-negative weighting changes both allocation and realized utility: R1-based routing has lower mean utility at every budget, and the selected sets are disjoint for $q \leq 20\%$. Practitioners can hold the evidence models fixed and report allocation disagreement and utility under each prespecified reward; choosing among contested objectives remains a governance decision. This retrospective analysis is not subgroup validation or prospective evidence.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main claims accurately, and the scope matches the theoretical and experimental results reported in Sections 1–3.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion discusses the fixed action/reward setup, calibration sensitivity, and constraints on realized workflow value. Appendices F.2 and G examine implementation and reward sensitivity.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The theoretical results and complete proofs appear in Section 2; Appendix A maps the claims to the Lean 4 proof artifacts.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper specifies datasets, data splits, reward matrices, cross-fitting procedure, baselines, and statistical tests sufficient to reproduce all main results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code, Lean 4 proofs, environment specifications, and reproduction/data-access instructions are available in the public repository linked on page 1.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Data splits, model configurations, reward regimes, baselines, and evaluation budgets are described in the experimental section and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Allocation comparisons report paired sign-test p-values with wins/losses/ties counts; the dissociation analysis reports mean Spearman correlations; sources of variability are stated throughout.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: AWS `g4dn.xlarge` (4 vCPUs, 16 GiB RAM; NVIDIA T4, 16 GB VRAM, 8.1 TFLOP/s FP32). Forward passes across the six CheXpert checkpoints (500 images each) and seven CIFAR-10 models (10,000 images each) require approximately 68 TFLOPs, counting two FLOPs per multiply–accumulate. At 20% model FLOP utilization, this corresponds to approximately 1 GPU-minute. CPU analyses are budgeted at minutes per primary run and 4 worker-hours for the full sweeps. No backbone training is performed.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses publicly available data and performs retrospective analysis; the authors have reviewed the NeurIPS Code of Ethics and identify no violations.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses potential benefits for clinical decision support and identifies risks from miscalibration or misuse in high-stakes deployment settings.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: No high-risk model or dataset is released.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Creators are cited and provider download/use agreements are followed. Licenses are identified where published: CIFAR-10H (CC BY-NC-SA 4.0), CheXpert images (Stanford University Dataset Research Use Agreement), TorchXRayVision/Lean 4/Mathlib code (Apache-2.0), and CIFAR model code (BSD-3-Clause). We release only our code and Lean 4 proofs, not source datasets or pretrained weights.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released experiment code and Lean 4 proofs are documented in the public repository, with data preparation and experiment instructions. We introduce no new dataset or pretrained model.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or human-subject experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human subjects; retrospective analysis of a public dataset.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not part of the core method.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.