DECISION-ALIGNED EVALUATION OF UNCERTAINTY QUANTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Uncertainty estimates in machine learning are typically evaluated using generic metrics such as the negative log-likelihood and expected calibration error, yet good performance on such metrics does not necessarily imply high utility in downstream decisions. We introduce *decision-alignment*, a criterion that reveals which metrics faithfully rank models according to their decision-making performance. Applying this framework, we show that many widely used uncertainty evaluation metrics are either misaligned with common decision problems or encode implausible prior beliefs about the downstream task. We then propose *prior-weighted utility metrics*, a special class of proper scoring rules that provides decision-aligned uncertainty evaluation. Across benchmark experiments and real-world case studies, our metrics consistently align with realized decision utility, while conventional metrics do not. Our results surface gaps in the current UQ evaluation protocol and offer a principled evaluation scheme toward decision-relevant UQ evaluation.

1 INTRODUCTION

Uncertainty quantification (UQ) in machine learning (ML) is key for reliable decision-making. Probabilistic ML research aims to develop high-quality UQ methods, facilitating ML usage in safety-critical domains. In this paper, we take *UQ evaluation* as our object of study. We focus on general-purpose evaluation, where the goal is to *rank models by performance* without a specific downstream task in mind—the typical setting of general ML research and model benchmarking.

Since the "true uncertainty" cannot be observed, general UQ evaluation relies on surrogate metrics such as negative log-likelihood (NLL) or expected calibration error (ECE) (Ovadia et al., 2019; Vadera et al., 2022; Basora et al., 2025). Each such metric ranks models according to a different property, e.g., the ECE favors models that are well-calibrated. However, ranking highly on such metrics does not necessarily imply high utility in downstream tasks (Vickers & Elkin, 2006; Dusenberry et al., 2020; Ferrer & Ramos, 2025)—even though reliable decision-making is usually the goal of UQ. Benchmarks guide progress in ML (Hardt & Recht, 2021). By restricting general-purpose UQ evaluation to generic measures such as the NLL and ECE, we risk leaving behind methods that do not excel on such metrics but are actually more useful in downstream applications, resulting in real financial damage for practitioners. We therefore argue that **UQ evaluation should be aligned with downstream decision utility**: models that perform well at the evaluation stage should also perform well on common downstream tasks.

In this work, we aim to improve alignment of UQ evaluation with decision-making. Our contribution is both theoretical and methodological. Concretely,

1. We introduce *decision-alignment* as a criterion to analyze UQ evaluation metrics in light of downstream decision-making. We find that most established metrics are not aligned with plausible decisions (section 3).
2. Based on these results, we propose *prior-weighted utility (PWU) metrics* that rank models according to their downstream decision-making performance (section 4).
3. We demonstrate in classification and regression experiments across several data modalities that only our PWU metrics reliably align with real downstream utilities (section 5).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

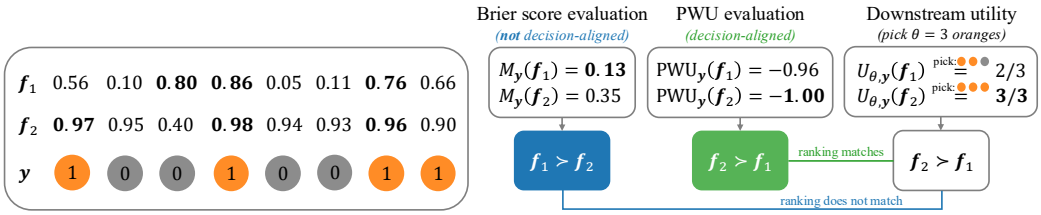


Figure 1: Example for our decision-alignment framework and PWU metrics. The box on the left shows two model predictions $\boldsymbol{f}_1, \boldsymbol{f}_2$ estimating the probability that a given orange is tasty (label 1), with corresponding true labels $\boldsymbol{y}$ below. The downstream decision-task is to select 3 tasty oranges (according to the 3 highest predicted probabilities in **bold**). Here, $\boldsymbol{f}_1$ leads to $2/3$ correct picks and $\boldsymbol{f}_2$ to $3/3$ correct picks, so $\boldsymbol{f}_2$ has the higher decision-utility. The boxes on the right show the evaluation stage: the Brier score (lower is better) ranks $\boldsymbol{f}_1$ first, consistent with the Brier score being **not decision-aligned**, while our **decision-aligned** PWU metric ranks $\boldsymbol{f}_2$ first.

2 Background & Related Work

Here, we briefly summarize our work's most relevant background. In appendix A, we detail the scope and positioning of our work. Further related work is acknowledged in appendix B.

Common UQ evaluation metrics in ML Current UQ studies in ML primarily evaluate uncertainty through predictive accuracy and calibration metrics, and do not explicitly reflect decision-quality. Benchmarking studies such as those of Ovadia et al. (2019), Nado et al. (2022), and Basora et al. (2025) typically assess models using NLL, Brier score, ECE, and ranking-scores such as the area under the receiver operating characteristic curve (AUROC). The benchmarking tools *URSABench* (Vadera et al., 2022) and *Neural Testbed* (Osband et al., 2022) do incorporate decision-making and bandit-style evaluations, respectively. However, these decision scenarios are artificial and simplified, which limits their transferability to the real world. In contrast, our PWU metrics are aligned with realistic downstream decisions in a principled manner.

Proper scoring rules Among the common UQ evaluation metrics, *proper scoring rules (PSRs)* such as the NLL and Brier score are often considered favorable as they are minimized at the ground-truth predictive distribution (Gneiting & Raftery, 2007; Brümmer, 2010; Dawid & Musio, 2014): If $f^* \in \mathbb{P}$ encodes the ground-truth law for outcome $y \in \mathbb{Y}$, then for a PSR $M : \mathbb{P} \times \mathbb{Y} \to \mathbb{R}$, it holds that

$$\mathbb{E}_{\mathrm{y}\sim f^*}\left[M(f^*, \mathrm{y})\right] \leq \mathbb{E}_{\mathrm{y}\sim f^*}\left[M(f, \mathrm{y})\right] \quad \text{for all} \quad f \in \mathbb{P},\ f \neq f^*$$

(if the inequality is strict, we say that M is *strictly* proper). That is, in expectation, every PSR would rank the ground-truth model first. In practice, however, the "true distribution" is not available, and **different PSRs do not necessarily rank models equally away from the minimizer**. Faithful UQ evaluation requires understanding what a metric's model ranking truly reflects, and our decision-alignment framework answers this question through the lens of decision-making.

3 Implicit Decision Beliefs of UQ Metrics

We begin by formalizing our decision-alignment framework. Afterwards, we use that framework to analyze popular UQ metrics, revealing what their model rankings reflect in terms of decision-making.

3.1 Decision-Alignment of UQ Metrics

Setup & notation We consider metrics $M : \mathbb{P}^n \times \mathbb{Y}^n \to \mathbb{R}$ and decision utilities $U_\theta : \mathbb{P}^n \times \mathbb{Y}^n \to \mathbb{R}$, parametrized by $\theta \in \Theta$, that evaluate a prediction-label pair $(\boldsymbol{f}, \boldsymbol{y})$, where $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathbb{Y}^n$ are i.i.d., and $\boldsymbol{f} = (f_1, \ldots, f_n) \in \mathbb{P}^n$ the corresponding predictions, $n \geq 2$. Our predictions are *probabilistic*, so each $f \in \mathbb{P}$ encodes a distribution over $\mathbb{Y}$. For example, in probabilistic binary classification, $\mathbb{Y} = \{0, 1\}$, and $f_i \in \mathbb{P} = [0, 1]$ is the predicted probability that datapoint i has label

$y_i = 1$. A common downstream task is top-k selection: pick k datapoints with label 1. Here, $\theta = k$, $\Theta = \mathbb{N}$, and its utility U_θ would be the fraction of correct picks (see figure 1 for an example). Its *Bayes-optimal* decision (termed *Bayes act*) is to pick the k instances with the highest expected utility assuming $\mathbf{y} \sim \boldsymbol{f}$. We restrict ourselves to prediction spaces $\mathbb{P}$ and utilities U_θ such that for each $\boldsymbol{f} \in \mathbb{P}^n$ a (not necessarily unique) *Bayes act* exists. Without loss of generality, we say that smaller values of M and larger values of U correspond to better performance. We use the notation[1]

$$M_{\boldsymbol{y}}(\boldsymbol{f}) := M(\boldsymbol{f}, \boldsymbol{y}) \quad \text{and} \quad U_{\theta,\boldsymbol{y}}(\boldsymbol{f}) := U_\theta(\boldsymbol{f}, \boldsymbol{y}).$$

Introducing decision-alignment We now formalize our framework for studying UQ evaluation metrics from a decision-theoretic perspective. Loosely speaking, a *decision-aligned* metric ranks models according to their performance on expected downstream tasks. Our framework is defined to encompass *all* metrics that satisfy this property, as we show in proposition 3.2.

Definition 3.1 (Decision-alignment)**.** A metric M is *decision-aligned* w.r.t. a decision family $\{U_\theta\}_{\theta\in\Theta}$ if there exists a *prior* $\pi : \Theta \to \mathbb{R}_{\geq 0}$ s.t. for any fixed labels $\boldsymbol{y}$, there exists a strictly increasing function $h_{\boldsymbol{y}} : \mathrm{Im}(M_{\boldsymbol{y}}) \to \mathbb{R}$ s.t.

$$h_{\boldsymbol{y}}\big(M_{\boldsymbol{y}}(\boldsymbol{f})\big) = \int_\Theta -U_{\theta,\boldsymbol{y}}(\boldsymbol{f})\,\pi(\theta)\,\mathrm{d}\theta \quad \forall \boldsymbol{f}. \tag{1}$$

The prior π can be thought of as a probability density, encoding a prior belief on the utility's parameter. For example, typical top-k selection tasks like recommender systems ask to select a *small* number k from a *large* data pool, so π could be designed to put more weight on smaller values of k. The prior of a decision-aligned metric describes what downstream task the metric implicitly expects. With this intuition in mind, the right-hand side of (1) reads as the *negative expected utility (NEU)* under $\theta \sim \pi$. Thus, we introduce the more intuitive notation

$$\mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}) := \int_\Theta -U_{\theta,\boldsymbol{y}}(\boldsymbol{f})\,\pi(\theta)\,\mathrm{d}\theta.$$

We now prove what we set out to formalize above: **a decision-aligned metric ranks models by their expected downstream performance** in the corresponding family U_θ, assuming $\theta \sim \pi$.

Proposition 3.2 (Decision-alignment $\Leftrightarrow$ strict order- and tie-preservation)**.** *Fix a decision family $\{U_\theta\}_{\theta\in\Theta}$ and a prior π. Then, for any fixed $\boldsymbol{y}$, the following statements are equivalent:*

(I) M is decision-aligned w.r.t. $\{U_\theta\}_{\theta\in\Theta}$ under π.

(II) M is strictly order-preserving and tie-preserving for $\{U_\theta\}_{\theta\in\Theta}$ under π, i.e., for all $\boldsymbol{f}_1, \boldsymbol{f}_2$,

$$M_{\boldsymbol{y}}(\boldsymbol{f}_1) < M_{\boldsymbol{y}}(\boldsymbol{f}_2) \Rightarrow \mathit{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}_1) < \mathit{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}_2),$$
$$M_{\boldsymbol{y}}(\boldsymbol{f}_1) = M_{\boldsymbol{y}}(\boldsymbol{f}_2) \Rightarrow \mathit{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}_1) = \mathit{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}_2).$$

Proof sketch. **(I)⇒(II):** Follows through strict monotonicity of $h_{\boldsymbol{y}}$. **(II)⇒(I):** Follows by defining $h_{\boldsymbol{y}}(m)$ as $\mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f})$ for any $\boldsymbol{f}$ with $M_{\boldsymbol{y}}(\boldsymbol{f}) = m$. We state the full proof in appendix C.1. □

Proposition 3.2 confirms the generality of our framework: $h_{\boldsymbol{y}}$ may be any strictly increasing map, so definition 3.1 includes *all* metrics that report the correct ranking, even if on a different scale.

Definition 3.1 also makes explicit that decision-alignment is a joint property of the triple (M, U_θ, π), not of M alone. Every PSR is decision-aligned with respect to *some* family and prior (see appendix D). Hence, decision-alignment can only be considered a useful property under reasonable U_θ and π.[2]

Remark 3.3 (On the nature of π). In our framework, a good intuition for π is a probability density, but formally, it can be any nonnegative and nonzero measure on Θ. It does not need to be finite or integrate to 1. We only require that the right-hand side of equation (1) is well-defined and finite. To simplify notation, we write $\pi(\theta)\mathrm{d}\theta$ for integration w.r.t. the measure π, including absolutely continuous priors with density $\pi(\theta)$, as well as purely atomic priors such as Dirac measures δ.

[1] Note that often, utility functions evaluate *actions* given labels, and we evaluate *predictions* given labels. To be precise, we define $U_{\theta,y}(f)$ as the utility of the Bayes act under f given y, so $U_{\theta,y}(f) := U_\theta(a^*(f), y)$, where $a^* := \arg\max_a \mathbb{E}_{\mathrm{y}\sim f} U_\theta(a, \mathrm{y})$. When the maximizer is not unique one can choose a tie-breaking rule.

[2] Note that the prior can only ever be identified up to a positive constant, because $(\pi, h_{\boldsymbol{y}})$ and $(a\pi, ah_{\boldsymbol{y}})$ satisfy equation (1) simultaneously for every $a > 0$. In appendix G, we prove that all implicit priors identified in our analysis are unique up to this scaling.

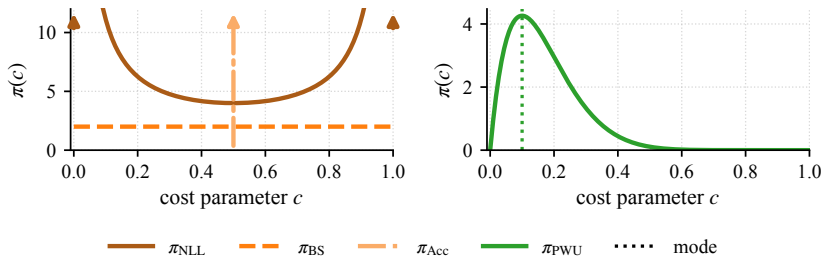


Figure 2: Implicit priors of common UQ metrics *(left)* and our PWU prior-choice *(right)* in the binary decision problem. The arrows indicate where the curves diverge. π_{PWU} reflects the asymmetric decision costs prevalent in many binary decisions, while the BS weighs each $c \in (0, 1)$ equally, $\pi_{\text{Acc}}(c)$ assumes perfectly symmetric costs, and $\pi_{\text{NLL}}(c)$ places unbounded mass on extreme regions.

Separability barrier Before we analyze common UQ metrics in terms of their decision-alignment, we introduce a lemma, which substantially simplifies our proofs. The lemma requires the notion of *pointwise-separability* and *coordinate-independence*.

Definition 3.4 (Pointwise-separability & coordinate-independence)**.** A metric/utility S is *pointwise-separable* if there exists a measurable *instance function* s s.t.

$$S(\boldsymbol{f}, \boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} s(f_i, y_i) \tag{2}$$

Otherwise, S is *non-separable*. S is *coordinate-independent* if the change in S through altering one prediction f_i does not depend on the other coordinates, that is, for $a, b \in \mathbb{P}$ and $\boldsymbol{r}, \boldsymbol{r}' \in \mathbb{P}^{n-1}$

$$S_{\boldsymbol{y}}((a, \boldsymbol{r})) < S_{\boldsymbol{y}}((b, \boldsymbol{r})) \Leftrightarrow S_{\boldsymbol{y}}((a, \boldsymbol{r}')) < S_{\boldsymbol{y}}((b, \boldsymbol{r}')),$$

where we used again $S_{\boldsymbol{y}}(\boldsymbol{f}) = S(\boldsymbol{f}, \boldsymbol{y})$, and analogously for equality and opposite strict inequality.[3] Otherwise, S is *coordinate-dependent*.[4]

Lemma 3.5 (Separability barrier)**.** *If the utility U_θ is pointwise-separable for every $\theta \in \Theta$ and $\theta \mapsto U_\theta(\boldsymbol{f}, \boldsymbol{y})$ is integrable for all $(\boldsymbol{f}, \boldsymbol{y})$, then any metric M that is decision-aligned w.r.t. $\{U_\theta\}_{\theta \in \Theta}$ (for some prior π) must be coordinate-independent.*

Proof sketch. By exchanging the integral and pointwise sum on the right-hand side of equation (1), we get that $h_{\boldsymbol{y}}\big(M_{\boldsymbol{y}}(\boldsymbol{f})\big)$ is pointwise-separable, which requires M to be coordinate-independent. We state the full proof in appendix C.2. □

Lemma 3.5 shows that coordinate-dependent metrics (e.g., calibration or ranking metrics) can immediately be ruled out as decision-aligned with respect to pointwise-separable utility families.[5]

Subsequently, we apply our framework to analyze several metric–utility pairs. We consider decision tasks that arise naturally in applications and span both pointwise-separable and non-separable utilities. Our analysis reveals that no metric we consider is aligned with plausible decisions.

3.2 Analysis of Classification Metrics

We first consider probabilistic binary classification. Model outputs are the predicted probabilities $f \in \mathbb{P} = [0, 1]$ that the true output label $y \in \mathbb{Y} = \{0, 1\}$ is 1. As evaluation metrics, we consider the

[3] We write $(f_i, \boldsymbol{r})$ for the prediction vector with value f_i at the i^{th} position and values $\boldsymbol{r}$ at the remaining $n - 1$ positions.

[4] Note that each pointwise-separable metric/utility is also coordinate-independent: For a pointwise-separable S, changing f_i only affects $s(f_i, y_i)$, which is independent of the remaining predictions.

[5] Coordinate-independence of M does not conversely imply separability of the utility family (see appendix E).

negative log-likelihood (NLL), Brier score (BS), accuracy (Acc), expected calibration error (ECE), maximum calibration error (MCE), area under the retention curve (R-AUC), and error detection (E-Det). We state the metric definitions in appendix K, where we also show that the **NLL, BS, and Acc are coordinate-independent**, and the **ECE, MCE, R-AUC, and E-Det are coordinate-dependent**. The results of this section are summarized on the top left of table F.1.

3.2.1 BINARY DECISION

We first consider a binary decision problem where an action in $\{0, 1\}$ (e.g., {do not treat, treat}) is taken based on the model's predicted probability $f \in [0, 1]$ for the positive class (e.g., disease present). We assume zero cost for correct actions and costs $C_{\text{FP}}, C_{\text{FN}} > 0$ for false positive and false negative actions (e.g., the false positive action would be to treat even though no disease is present). Without loss of generality, we normalize the costs as $c_{\text{FP}} = C_{\text{FP}}/(C_{\text{FP}} + C_{\text{FN}})$ (so $c_{\text{FN}} = 1 - c_{\text{FP}}$). For ease of notation, we write $c = c_{\text{FP}}$ in the following. With this setup, the Bayes act is to select action 1 whenever $f > c$ (Elkan, 2001). This allows us to write the pointwise utility of a prediction f with true label $y \in \mathbb{Y}$ for a fixed cost $c \in (0, 1)$ as

$$u_c(f, y) = -y(1 - c)\mathbf{1}_{f \leq c} - (1 - y)c\mathbf{1}_{f > c}.$$

The dataset-level utility $U_c(\boldsymbol{f}, \boldsymbol{y})$ is defined as the average pointwise utility. As a direct consequence, $\boldsymbol{U_c}$ **is pointwise-separable**. By lemma 3.5, only the NLL, BS, and Acc can be decision-aligned for $\{U_c\}_{c\in(0,1)}$. For the specific downstream utility u_c, it is already known that PSRs l admit an integral representation—a result that inspired our framework. Under mild conditions (Savage, 1971; Schervish, 1989; Buja et al., 2005; Gneiting & Raftery, 2007; Brümmer, 2010),

$$l(f, y) = \int_0^1 -u_c(f, y)\, \pi(c)\, \mathrm{d}c. \tag{3}$$

The expressions for $\pi(c)$ for the NLL, BS, and Acc have been stated before (see, e.g., Buja et al., 2005; Gneiting & Raftery, 2007).

Proposition 3.6 (Buja et al. (2005))**.** *The NLL, BS, and Acc are decision-aligned w.r.t.* $\{U_c\}_{c\in(0,1)}$ *with* $h_{\boldsymbol{y}} = \mathrm{id}$ *for all* $\boldsymbol{y}$,[6] *and*

$$\pi_{NLL}(c) = 1/(c(1 - c)),\ \pi_{BS} = 2,\ \pi_{Acc}(c) = 2\delta_{0.5}(c).$$

These priors are unique up to a positive constant (proposition G.4), so they are *the* decision beliefs encoded by the NLL, BS, and Acc. The decision-theoretic interpretation (3) suggests that good performance on PSRs implies good decision-making. However, this conclusion is refuted once we audit whether the priors from proposition 3.6 (visualized on the left of figure 2) actually represent plausible assumptions about the downstream task. In many binary decisions, the cost of a false negative (e.g., missed disease) is much higher than the cost of a false positive (e.g., unnecessary treatment) (Elkan, 2001). In contrast, the BS weighs each $c \in (0, 1)$ equally, $\pi_{\text{Acc}}(c)$ assumes that the costs are perfectly symmetric, and $\pi_{\text{NLL}}(c)$ places unbounded mass on extreme regions, rewarding performance where one error type is essentially free and a trivial policy suffices. So while these metrics do measure relevant properties like 'how well does the predictive distribution anticipate the data?', their implicit priors are *implausible*.[7] Consequently, their model ranking in terms of a downstream binary decision is misleading, exposing the need for new metrics with plausible priors.

Since the ECE, MCE, R-AUC, and E-Det are coordinate-dependent, it follows directly from lemma 3.5 that they are not decision-aligned w.r.t. $\{U_c\}_{c\in(0,1)}$.

Corollary 3.7. *The ECE, MCE, R-AUC, and E-Det are not decision-aligned w.r.t.* $\{U_c\}_{c\in(0,1)}$.

3.2.2 TOP-k SELECTION

Next, we consider the task we employed as an introductory example in section 3.1: Top-k selection for $k \in [n-1] := \{1, \ldots, n-1\}$, which arises, for instance, in the context of recommender systems.

[6] With our lower-is-better convention for Acc (appendix K), we get $h_{\boldsymbol{y}}(m) = m + 1$; omitted for simplicity.

[7] Strictly speaking, whether a prior is plausible or not depends on a given decision context and cannot be posed as a general property. We use the term *implausible* to flag priors that most practitioners would not explicitly choose, and our experiments support our judgment. However, we ask the reader to read our implausibility tag as an intuitive, but *not* as a formal, globally true property.

As downstream utility, we use the established *precision@k* metric (see, e.g., Tamm et al., 2021),

$$U_k(\boldsymbol{f}, \boldsymbol{y}) = \frac{1}{k} \sum_{i \in \mathbb{S}_k(\boldsymbol{f})} y_i,$$

where $\mathbb{S}_k(\boldsymbol{f})$ is the index set of the selected instances according to the predicted scores.[8] The Bayes act is to select the k instances with the highest scores (Wydmuch et al., 2018). **U_k is non-separable**: whether an instance contributes to U_k depends on its rank relative to all other predictions, so U_k cannot be written as a sum of independent terms. Hence, we cannot leverage lemma 3.5. Nevertheless, we can prove that no metric we consider is decision-aligned w.r.t. $\{U_k\}_{k \in [n-1]}$.

Proposition 3.8. *The NLL, BS, Acc, ECE, MCE, R-AUC, and E-Det are not decision-aligned w.r.t.* $\{U_k\}_{k \in [n-1]}$.

Proof sketch. The value of U_k is only influenced by the subset $\mathbb{S}_k(\boldsymbol{f})$, and all metrics we consider evaluate on the entire dataset. One can find predictions $\boldsymbol{f}_1, \boldsymbol{f}_2$ that yield different metric values but leave the expected utility unchanged. We state the full proof in appendix C.3. □

We present our remaining decision-alignment analysis in appendix F, covering multiclass classification and regression (uni- and multivariate). In total, we analyze 32 metric–utility pairs, resulting in a comprehensive guide for how to interpret model rankings in terms of decision-making, summarized in table F.1. Our analysis can be extended to new metric–utility pairs of interest in the same manner.

4 Prior-weighted utility metrics

The key takeaway from section 3 is that common UQ evaluation metrics either encode implausible decision beliefs or do not reflect common decision utility at all. In other words, while established metrics report valuable information (e.g., whether a model is well-calibrated), **current UQ benchmarks silently optimize for the wrong downstream world**, potentially resulting in real financial damage for downstream decision-makers. To close this gap in the UQ evaluation protocol, we propose *prior-weighted utility (PWU)* metrics. By definition, they are *decision-aligned* and encode *plausible prior beliefs* about downstream decisions.[9]

4.1 PWU construction

To define a PWU metric, we select a decision family $\{U_\theta\}_{\theta \in \Theta}$ and a plausible prior π such that $\mathrm{NEU}_{\theta \sim \pi, \boldsymbol{y}}(\boldsymbol{f}) < \infty$ for all $(\boldsymbol{f}, \boldsymbol{y})$. To align the PWU metric with $\{U_\theta\}_{\theta \in \Theta}$, we apply equation (1) directly with $h_{\boldsymbol{y}} = \mathrm{id}$ for all $\boldsymbol{y}$:

$$\mathrm{PWU}(\boldsymbol{f}, \boldsymbol{y}) := \int_\Theta -U_\theta(\boldsymbol{f}, \boldsymbol{y})\, \pi(\theta)\, \mathrm{d}\theta. \tag{4}$$

We propose a PWU metric for each decision problem we analyzed. For example, for binary classification, we obtain one PWU metric each for the binary decision (BD) and top-k selection (TS) problems,

$$\text{PWU-BD}(\boldsymbol{f}, \boldsymbol{y}) = \int_0^1 -U_c(\boldsymbol{f}, \boldsymbol{y})\, \pi_c(c)\, \mathrm{d}c, \quad \text{PWU-TS}(\boldsymbol{f}, \boldsymbol{y}) = \sum_{k=1}^{n-1} -U_k(\boldsymbol{f}, \boldsymbol{y})\, \pi_k(k),$$

where U_c, U_k are defined as in section 3.2, and we carefully choose corresponding *plausible* priors π_c, π_k, visualized in figure I.1 (π_c also on the right of figure 2). In appendix I, we justify them explicitly and provide some general guidance on prior elicitation. Note that constructing a PWU metric is a general method that can be applied to a broader set of decisions (and priors) than just those considered in this paper. Since each metric corresponds to a single decision family, it is advisable to use a *variety* of PWU metrics in UQ benchmarking to test models across different decision contexts.

[8] For $\mathbb{S}_k(\boldsymbol{f})$ to be uniquely defined, one can choose a tie-breaking rule.

[9] Appendix H states a structured PWU documentation, similar to *model cards* (Mitchell et al., 2019).

4.2 METRIC PROPERTIES

The purpose of our PWU metrics is to evaluate UQ in terms of downstream decision utility. Thus, we intentionally defined them to be decision-aligned with their corresponding utility family.

Proposition 4.1 (PWU metrics are decision-aligned)**.** *Pick any decision family $\{U_\theta\}_{\theta\in\Theta}$ and a prior π over Θ. Then, the resulting PWU metric (4) is decision-aligned w.r.t. $\{U_\theta\}_{\theta\in\Theta}$ under π.*

Proof. Follows by comparing the PWU definition (4) to definition 3.1 with $h_{\boldsymbol{y}} = \mathrm{id}$ for all $\boldsymbol{y}$. □

Recall that decision-alignment is equivalent to strict order- and tie-preservation. Hence, PWU metrics offer a principled and decision-theoretic value for UQ benchmarking. Furthermore, using a PWU metric does not sacrifice properness—all PWU metrics are proper, as we show next.

Proposition 4.2 (PWU metrics are proper)**.** *Pick any decision family $\{U_\theta\}_{\theta\in\Theta}$ and a corresponding prior π, let $\boldsymbol{f}^* \in \mathbb{P}^n$ denote the ground truth, and assume $\int_\Theta \mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}\left[|U_{\theta,\mathbf{y}}(\boldsymbol{f})|\right]\,\pi(\theta)\,\mathrm{d}\theta < \infty$ for all $\boldsymbol{f} \in \mathbb{P}^n$. Then, the resulting PWU metric (4) is a PSR, and it is strictly proper if $\{U_\theta\}_{\theta\in\Theta}$ separates predictions under π (condition (8) in proposition D.1).*

Proof. The PWU (4) is decision-aligned by proposition 4.1 and, by comparing (4) to definition 3.1, we have $h_{\boldsymbol{y}} = \mathrm{id}$ for all $\boldsymbol{y}$. Applying proposition D.1 with $a = 1, b = 0$, we can conclude that the PWU metric is proper (and strictly proper under condition (8) in proposition D.1). □

Our PWUs do not come without limitations (see also appendix H). No single PWU can cover all downstream tasks, so we recommend evaluating on a *variety* of PWUs. Furthermore, PWU priors may be misspecified. In appendix J.3 we show that **PWU metrics retain positive utility-alignment even under severe prior perturbation**, confirming PWUs' suitability for general UQ benchmarking.

5 EXPERIMENTS

We now verify empirically that our PWU metrics align best with common downstream decision problems and show that this has real practical consequences: deploying the best model according to PWU metrics leads to substantially lower decision regret in real-world case studies. Our anonymized code is provided as supplementary material and will be made public upon acceptance.

5.1 CONTROLLED EXPERIMENTS ON BENCHMARK DATASETS

We start by measuring metric alignment on a large benchmark suite. For this, we deploy our metrics to score the performance of 70 probabilistic models in total; fitting each on 5 datasets per prediction task and data modality (data: tabular, image, text; 40 datasets in total; tasks: binary/multiclass classification and uni-/multivariate regression). We then score the same models on all utilities defined in this paper (e.g., binary decision and top-k selection; 6 downstream tasks in total). We compare each metric's model ranking to each applicable utility-ranking. Standard measures to quantify ranking similarity are Spearman's ρ and Kendall's τ, and we use Kendall's τ because it is robust and interpretable for small rankings. For example, if a metric exhibits $\tau = 0.8$ on some downstream task, it means that in a pairwise comparison of the metric's ranking to the downstream utility's ranking, 90% of the pairs agree. We repeat the τ-measurement on 100 differently sampled test sets to quantify the results' variability. We intentionally use datasets and models of different types, sizes, dimensionality, and complexity to achieve representative results. For further details on our experimental setup, protocol, models, and datasets, we refer to appendix J.

In figure 3, we report our binary classification results (remaining results in figures J.1 and J.2). Each panel summarizes the results across 5 datasets, and visualizes the ranking similarity when scoring 10 different models. We find that even though some conventional metrics are decision-aligned, their implicit priors negatively impact their utility-alignment. This supports our judgment that decision-alignment alone is not sufficient: a *plausible* prior is what makes decision-alignment a beneficial property. Not surprisingly, metrics that are not decision-aligned usually do not exhibit strong alignment. In contrast, our PWU metrics align best with their corresponding utility family, and this holds consistently across tabular, image, and text data.

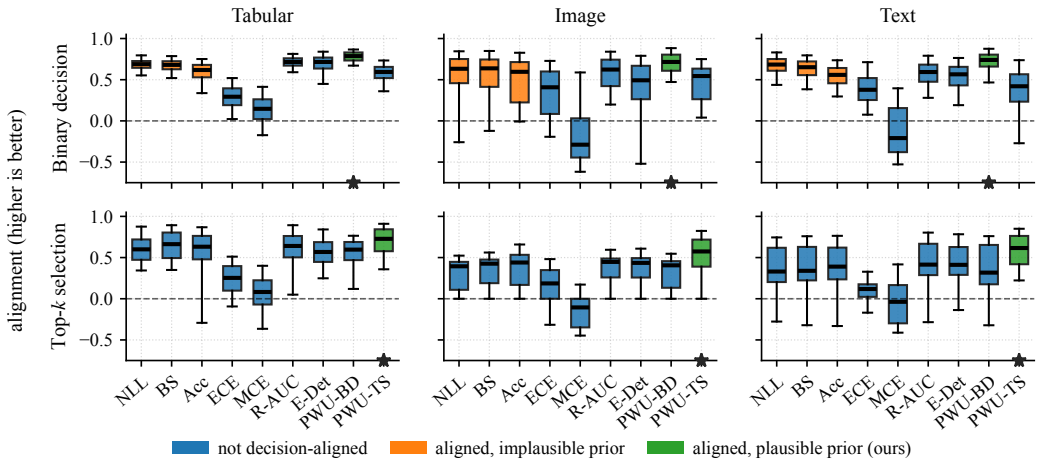


Figure 3: Metric–utility alignment (Kendall's τ) in our binary classification experiments. Each box's black line shows the median Kendall's τ value, the box spans the 25$^{\text{th}}$-75$^{\text{th}}$ percentile, and the whiskers the 5$^{\text{th}}$-95$^{\text{th}}$ percentile over 100 repeats; averaged over five datasets per modality. The coloring corresponds to our decision-alignment analysis. We mark the metric with the highest median with a star. Our PWU metrics align best with their corresponding utility family on every data modality.

5.2 Applied Case Studies

While the above experiments consider the same utilities as our PWUs, we now deviate from this setup and implement three applied case studies. In our first case study, we consider a wind farm operator who wants to sell electricity in the *day-ahead market*. The action is to place bids one day in advance. If the actual wind power production deviates from the bid, they risk a penalty for the resulting shortfall or surplus. We follow the setup from Bruninx et al. (2025), who derive an optimal bidding policy under a risk budget: the operator bids more aggressively when the forecast is more certain, and falls back to the point forecast when uncertainty is too high. The utility (payoff) is market-dependent for every bid and unknown in advance, deviating significantly from the setup any of our PWUs anticipates. We explain the case study in more detail in appendix J.2 and report the alignment results in table 1 and on the left of figure J.10. In this real-life setup, all conventional UQ metrics align only weakly with bidding performance, again showcasing how disconnected generic metrics are from real-world tasks. In contrast, our two PWU metrics[10] exhibit the strongest positive utility alignment.

Table 1: Metric–utility alignment (Kendall's τ) in the electricity market case study. We report the median and the 5th and 95th percentile. We mark the highest median in **bold** and underline the 2$^{\text{nd}}$-highest. Our PWU metrics have the strongest bidding-utility alignment.

NLL	MSE	ECE	MCE	R-AUC	E-Det	PWU-SP	PWU-TS
0.11	0.27	0.27	0.21	0.07	0.07	**0.47**	0.29
[-0.20, 0.47]	[-0.16, 0.69]	[-0.16, 0.60]	[-0.12, 0.56]	[-0.34, 0.47]	[-0.42, 0.42]	[-0.02, 0.73]	[-0.07, 0.56]

In our remaining two case studies (see appendix J.2), we consider the tasks of credit approval and peer-to-peer lending, again using real-world economic payoffs as utility functions. We observe once more that our PWU metrics align well with the decision utilities—even though some decision parameters differ strongly from our chosen PWU prior (see figure J.9). This demonstrates that PWU metrics remain robust under prior misspecification, a finding further supported by our systematic sensitivity analysis (see appendix J.3).

[10]Note that this study is a regression task and hence comes with different PWUs than defined in the main text. See appendix H for the full list of our PWUs.

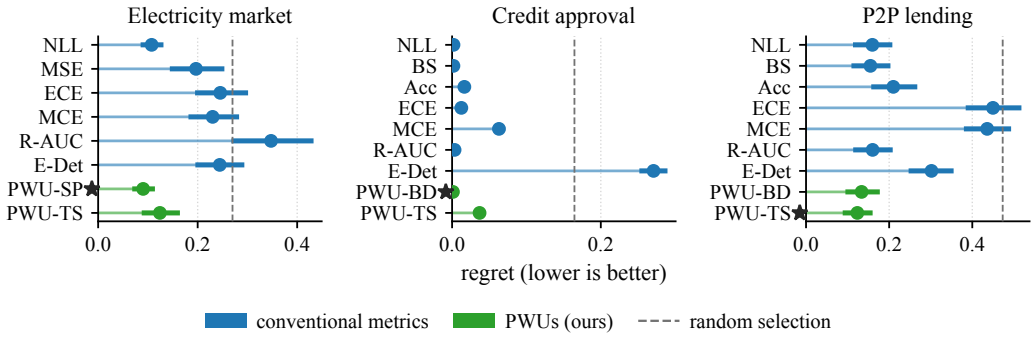


Figure 4: Mean realized regret of model deployment as fraction of the achievable utility range (0 = always picks the best model), averaged over 100 repeats. The dot is the mean, and the bar a percentile bootstrap 95% confidence interval. We mark the metric with the lowest mean with a star. In each case study, one of our PWU metrics achieves lowest mean regret.

5.3 Regret analysis

The results so far confirm that alignment holds in practice, both in a controlled setup and outside of our PWUs' anticipated regime. We now show that this positive alignment result has real practical consequences: benchmarks' model rankings decide which models are further developed and eventually deployed by practitioners. If we pursue models in probabilistic ML research solely based on common UQ metrics, downstream practitioners will face higher losses. To capture this phenomenon, we revisit our case studies (section 5.2) and measure each metric's *realized decision-regret*, that is, the downstream utility lost when deploying the metric's top-ranked model instead of the utility-optimal one. In figure 4, we can see that our PWU metrics attain the lowest decision-regret in all three case studies. Most strikingly, in the electricity market case study, selecting by R-AUC forfeits on average 35% of the achievable bidding-profit range, vs. only 9% for our PWU-SP—a loss of €250k vs. €66k in original units. The same analysis across our benchmark experiments (section 5.1) shows our PWU metrics consistently attaining low decision regret (see figures J.3, J.4, and J.5).

5.4 Ablations

Our PWUs' priors may be misspecified in light of certain decision problems. To analyze whether this impairs our PWUs' performance, we test their sensitivity to several degrees of prior misspecification in appendix J.3. We find that our PWUs' performance remains stable across all levels of misspecification. We also test our findings' robustness against the choice of our alignment measure (Kendall's τ). We report results from alternative alignment measures (top-1 and top-3 agreement, Pearson's correlation) in figures J.6, J.7, and J.8, and find that our PWUs perform equally well.

6 Conclusion

We introduced decision-alignment as a criterion that connects uncertainty evaluation metrics to downstream decision utility and used it to diagnose that many standard UQ metrics are misaligned with common decision problems. Our results suggest that current UQ evaluation can be systematically misleading when the goal is to support downstream decisions. Motivated by this insight, we proposed PWU metrics, a principled class of decision-aligned PSRs that enable utility-relevant benchmarking. Across benchmark datasets and real-world case studies, PWU metrics consistently produce model rankings that better reflect realized decision utility than conventional UQ metrics. For future general-purpose UQ evaluation, we recommend reporting PWU metrics for at least two distinct decision families that plausibly reflect deployment use cases, stating the prior π explicitly, and reporting conventional metrics alongside for their diagnostic signal (e.g., whether a model is well-calibrated). This extension requires only a modest amount of additional reporting, but it shifts UQ benchmarking towards bringing methods to light that enable reliable decision-making.

AI USE STATEMENT

In this work, we used generative AI tools for helping develop theoretical models or conceptual frameworks, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting datasets, and supporting qualitative and thematic data analysis. We have not used generative AI tools for formulating mathematical claims, proposing or refining hypotheses, and interpreting results, and generating synthetic datasets is not applicable to this work. Additionally, we used generative AI tools for creating or modifying scientific figures or images, suggesting experimental parameters, creating or editing software code, drafting parts of the appendix, discovering research topics or identifying gaps, brainstorming, sourcing/searching for information, editing a research paper to improve readability, and identifying relevant literature. We have reviewed all AI-assisted work. Each mathematical proof was proofread and corrected in detail by two authors, and LLM-generated code was validated by one author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work concerns evaluation methodology and relies exclusively on publicly available benchmark datasets, involving no human subjects or newly collected data. The primary use case of our method is general-purpose model benchmarking, which raises no immediate ethical concerns. If PWU metrics are used by practitioners in sensitive domains (e.g., credit approval), we note that results depend on the specific utility function, and practitioners should ensure that the utilities and priors they encode reflect legitimate and fair cost decisions.

REPRODUCIBILITY STATEMENT

Our anonymized codebase is provided as supplementary material and will be deanonymized upon acceptance. The codebase provides a `README.md` with clear instructions on how to run the code and reproduce our results. All randomness is controlled through fixed seeds. For all theoretical claims, we provide a formal proof in this manuscript. All datasets are publicly available and their preprocessing code, directly starting at the raw files, is provided in our codebase.

REFERENCES

Stefan Aeberhard and M Forina. Wine, 1992. URL https://doi.org/10.24432/C5PC7J.

E Alpaydin and Fevzi Alimoglu. Pen-Based Recognition of Handwritten Digits, 1996. URL https://doi.org/10.24432/C5MG6K.

Alejandro Correa Bahnsen, Djamia Aouada, and Bjorn Ottersten. Example-Dependent Cost-Sensitive Logistic Regression for Credit Scoring. In *13th International Conference on Machine Learning and Applications*, pp. 263–269, Detroit, MI, USA, 2014. IEEE. URL http://ieeexplore.ieee.org/document/7033125/.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1644–1650, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.findings-emnlp.148/.

Luis Basora, Arthur Viens, Manuel Arias Chao, and Xavier Olive. A benchmark on uncertainty quantification for deep learning prognostics. *Reliability Engineering & System Safety*, 253:110513, 2025. URL https://linkinghub.elsevier.com/retrieve/pii/S0951832024005854.

Barry Becker and Ronny Kohavi. Adult, 1996. Published: UCI Machine Learning Repository.

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Pitfalls of Epistemic Uncertainty Quantification through Loss Minimisation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 29205–29216. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/bc1d640f841f752c689aae20b31198c1-Paper-Conference.pdf.

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. On Second-Order Scoring Rules for Epistemic Uncertainty Quantification. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 2078–2091. PMLR, July 2023. URL https://proceedings.mlr.press/v202/bengs23a.html.

J Blackard. Covertype, 1998. URL https://doi.org/10.24432/C50K5N.

R. Bock. MAGIC Gamma Telescope, 2004. Published: UCI Machine Learning Repository.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, pp. 491–500, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 978-1-4503-6675-5. doi: 10.1145/3308560.3317593. URL https://doi.org/10.1145/3308560.3317593.

Victor Bouvier, Simona Maggio, Alexandre Abraham, and Léo Dreyfus-Schmidt. Towards Clear Expectations for Uncertainty Estimation, 2022. URL http://arxiv.org/abs/2207.13341.

A. Allen Bradley and Stuart S. Schwartz. Summary Verification Measures and Their Interpretation for Ensemble Forecasts. *Monthly Weather Review*, 139(9):3075–3089, 2011. URL https://journals.ametsoc.org/doi/10.1175/2010MWR3305.1.

Gary Bradshaw. Solar Flare, 1989. URL https://doi.org/10.24432/C5530G.

Max Bruninx, Timothy Verstraeten, Jalal Kazempour, and Jan Helsen. Day-Ahead Bidding Strategies for Wind Farm Operators Under a One-Price Balancing Scheme. In *Proceedings of the 16th ACM International Conference on Future and Sustainable Energy Systems*, pp. 719–726, Rotterdam Netherlands, 2025. ACM. URL https://dl.acm.org/doi/10.1145/3679240.3734639.

Niko Brümmer. *Measuring, Refining and Calibrating Speaker and Language Information Extracted From Speech*. PhD Thesis, Stellenbosch University, 2010. URL https://scispace.com/pdf/measuring-refining-and-calibrating-speaker-and-language-53glvzuiuj.pdf.

Sven Buechel and Udo Hahn. EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In Mirella Lapata, Phil Blunsom, and Alexander Koller (eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 578–585, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2092/.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications, 2005. URL http://stat.wharton.upenn.edu/~buja/PAPERS/paper-proper-scoring.pdf.

Ajay Byanjankar, József Mezei, and Markku Heikkilä. Data-Driven Optimization of Peer-to-Peer Lending Portfolios Based on the Expected Value Framework. *Intelligent Systems in Accounting, Finance and Management*, 28(2):119–129, 2021. URL https://onlinelibrary.wiley.com/doi/10.1002/isaf.1490.

P Cortez, A Cerdeira, F Almeida, T Matos, and J Reis. Wine Quality, 2009. URL https://doi.org/10.24432/C56S3T.

Peng Cui, Wenbo Hu, and Jun Zhu. Calibrated Reliable Regression using Maximum Mean Discrepancy. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 33 of *NIPS '20*, pp. 17164–17175, Vancouver, BC, Canada, 2020. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c74c4bf0dad9cbae3d80faa054b7d8ca-Paper.pdf.

Will Cukierski. GiveMeSomeCredit, 2011. URL https://kaggle.com/competitions/GiveMeSomeCredit. Credit Fusion.

Etienne David, Simon Madec, Pouria Sadeghi-Tehran, Helge Aasen, Bangyou Zheng, Shouyang Liu, Norbert Kirchgessner, Goro Ishikawa, Koichi Nagasawa, Minhajul A. Badhon, Curtis Pozniak, Benoit de Solan, Andreas Hund, Scott C. Chapman, Frédéric Baret, Ian Stavness, and Wei Guo. Global Wheat Head Detection (GWHD) Dataset: A Large and Diverse Dataset of High-Resolution RGB-Labelled Images to Develop and Benchmark Wheat Head Detection Methods. *Plant Phenomics*, 2020:3521852, 2020. URL https://doi.org/10.34133/2020/3521852.

Alexander Philip Dawid and Monica Musio. Theory and applications of proper scoring rules. *METRON*, 72(2):169–183, 2014. URL http://link.springer.com/10.1007/s40300-014-0039-y.

Rabanus Derr and Robert C. Williamson. Forecast Evaluation and the Relationship of Regret and Calibration, 2025. URL https://arxiv.org/abs/2401.14483v3.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. URL https://aclanthology.org/N19-1423/.

Tony Duan, Anand Avati, Daisy Yi Ding, Khanh K Thai, Sanjay Basu, Andrew Ng, and Alejandro Schuler. NGBoost: Natural Gradient Boosting for Probabilistic Prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020. URL https://proceedings.mlr.press/v119/duan20a.html.

Michael W. Dusenberry, Dustin Tran, Edward Choi, Jonas Kemp, Jeremy Nixon, Ghassen Jerfel, Katherine Heller, and Andrew M. Dai. Analyzing the role of model uncertainty for electronic health records. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 204–213, Toronto Ontario Canada, 2020. ACM. URL https://dl.acm.org/doi/10.1145/3368555.3384457.

Werner Ehm, Tilmann Gneiting, Alexander Jordan, and Fabian Krüger. Of Quantiles and Expectiles: Consistent Scoring Functions, Choquet Representations and Forecast Rankings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(3):505–562, 2016. URL https://academic.oup.com/jrsssb/article/78/3/505/7040984.

Elia. Elia Open Data Platform, 2025. URL https://opendata.elia.be.

Charles Elkan. The Foundations of Cost-Sensitive Learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, volume 2 of *IJCAI'01*, pp. 973–978, Seattle, WA, USA, 2001. Morgan Kaufmann Publishers Inc. URL https://dl.acm.org/doi/abs/10.5555/1642194.1642224.

European Network of Transmission System Operators for Electricity ENTSO-E. ENTSO-E Transparency Platform, 2025. URL https://transparency.entsoe.eu.

Manaal Faruqui and Dipanjan Das. Identifying Well-formed Natural Language Questions. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 798–803, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1091. URL https://aclanthology.org/D18-1091/.

Luciana Ferrer and Daniel Ramos. Evaluating Posterior Probabilities: Decision Theory, Proper Scoring Rules, and Calibration. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=qbrE0LR7fF.

R Fisher. Iris, 1936. URL https://doi.org/10.24432/C56C76.

Gerardo Flores, Abigail Schiff, Alyssa H. Smith, Julia A. Fukuyama, and Ashia C. Wilson. A Consequentialist Critique of Binary Classification Evaluation: Theory, Practice, and Tools, 2026. URL http://arxiv.org/abs/2504.04528.

Gerardo A Flores, Alyssa H Smith, Julia A Fukuyama, and Ashia C Wilson. Aligning Evaluation with Clinical Priorities: Calibration, Label Shift, and Error Costs. In *The 39th Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=b7Ka31HvXU.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, New York, New York, USA, June 2016. PMLR. URL https://proceedings.mlr.press/v48/gal16.html.

Jacob R Gardner, Geoff Pleiss, David Bindel, Kilian Q Weinberger, and Andrew Gordon Wilson. GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. In *Advances in Neural Information Processing Systems*, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/27e8e17134dd7083b050476733207ea1-Abstract.html.

Nathan George. Lending Club Loan Data, 2020. URL https://www.kaggle.com/datasets/wordsforthewise/lending-club.

Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. URL http://www.tandfonline.com/doi/abs/10.1198/016214506000001437.

Tilmann Gneiting and Roopesh Ranjan. Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules. *Journal of Business & Economic Statistics*, 29(3):411–422, 2011. ISSN 0735-0015, 1537-2707. doi: 10.1198/jbes.2010.08110. URL http://www.tandfonline.com/doi/abs/10.1198/jbes.2010.08110.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting Deep Learning Models for Tabular Data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18932–18943. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/9d86d83f925f2149e9edb0ac3b49229c-Paper.pdf.

Sebastian G. Gruber and Florian Buettner. Better Uncertainty Calibration via Proper Scores for Classification and Beyond. In *36th Conference on Neural Information Processing Systems*, 2022. URL http://arxiv.org/abs/2203.07835.

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*, volume 70, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Moritz Hardt and Benjamin Recht. Patterns, predictions, and actions: A story about machine learning, 2021. URL http://arxiv.org/abs/2102.05242.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Los Alamitos, CA, USA, 2016. IEEE Computer Society. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.90.

Mahmoud Hegazy, Michael I. Jordan, and Aymeric Dieuleveut. Scalable Utility-Aware Multiclass Calibration, 2025. URL http://arxiv.org/abs/2510.25458.

James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable Variational Gaussian Process Classification. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38, pp. 351–360. PMLR, 2015. URL https://proceedings.mlr.press/v38/hensman15.html.

Jose Hernandez-Orallo, Peter Flach, and Cesar Ferri. A Unified View of Performance Metrics: Translating Threshold Choice into Expected Classification Loss. *Journal of Machine Learning Research*, 13:2813–2869, 2012.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. In *The 11th International Conference on Learning Representations*, 2023. URL https://openreview.net/pdf?id=cp5PvcI6w8_.

Lunjia Hu and Yifan Wu. Calibration Error for Decision Making, 2024. URL http://arxiv.org/abs/2404.13503.

Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The UCI Machine Learning Repository, 2025. URL https://archive.ics.uci.edu.

Robert Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. U-Calibration: Forecasting for an Unknown Agent. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 5143–5145. PMLR, 2023. URL https://proceedings.mlr.press/v195/kleinberg23a.html.

M Koklu and Ilker Ali Özkan. Dry Bean, 2020. URL https://doi.org/10.24432/C50S4B.

Alex Krizhevsky and others. Learning multiple layers of features from tiny images, 2009. URL https://cave.cs.toronto.edu/kriz/learning-features-2009-TR.pdf.

Marek Kuczma and Attila Gilányi (eds.). *An Introduction to the Theory of Functional Equations and Inequalities: Cauchy's Equation and Jensen's Inequality*. SpringerLink Bücher. Birkhäuser Basel, Basel, second edition, 2009. URL https://doi.org/10.1007/978-3-7643-8749-5.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

Max-Heinrich Laves, Sontje Ihler, Jacob F Fast, Luder A Kahrs, and Tobias Ortmaier. Well-Calibrated Regression Uncertainty in Medical Imaging with Deep Learning. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, volume 121, pp. 393–412. PMLR, 2020. URL https://proceedings.mlr.press/v121/laves20a.html.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019. URL https://arxiv.org/abs/1907.11692.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning Word Vectors for Sentiment Analysis. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015/.

Andrey Malinin, Neil Band, Alexander Ganshin, German Chesnokov, Mark J. F. Gales, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, Vyas Raina, Denis Roginskiy, Mariya Shmatova, Panos Tigas, and Boris Yangel. Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks. In *35th Conference on Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://openreview.net/forum?id=qM45LHaWM6E.

Shireen Kudukkil Manchingal, Andrew Bradley, Julian F. P. Kooij, Keivan Shariatmadar, Neil Yorke-Smith, and Fabio Cuzzolin. Epistemic Artificial Intelligence is Essential for Machine Learning Models to Truly'Know When They Do Not Know', 2025a. URL https://arxiv.org/abs/2505.04950.

Shireen Kudukkil Manchingal, Muhammad Mubashar, Kaizheng Wang, Keivan Shariatmadar, and Fabio Cuzzolin. Random-Set Neural Networks. In *International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=pdjkikvCch.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dSprites: Disentanglement testing Sprites dataset, 2017. URL https://github.com/deepmind/dsprites-dataset/.

Petrus Mikkola, Osvaldo A. Martin, Suyog Chandramouli, Marcelo Hartmann, Oriol Abril Pla, Owen Thomas, Henri Pesonen, Jukka Corander, Aki Vehtari, Samuel Kaski, Paul-Christian Bürkner, and Arto Klami. Prior knowledge elicitation: The past, present, and future. *Bayesian Analysis*, 19(4): 1129–1161, 2024. URL https://doi.org/10.1214/23-BA1381.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 220–229, New York, NY, USA, 2019. Association for Computing Machinery. URL https://doi.org/10.1145/3287560.3287596.

S Moro, P Rita, and P Cortez. Bank Marketing, 2014. URL https://doi.org/10.24432/C5K306.

Zachary Nado, Neil Band, Mark Collier, Josip Djolonga, Michael W. Dusenberry, Sebastian Farquhar, Qixuan Feng, Angelos Filos, Marton Havasi, Rodolphe Jenatton, Ghassen Jerfel, Jeremiah Liu, Zelda Mariet, Jeremy Nixon, Shreyas Padhy, Jie Ren, Tim G. J. Rudner, Faris Sbahi, Yeming Wen, Florian Wenzel, Kevin Murphy, D. Sculley, Balaji Lakshminarayanan, Jasper Snoek, Yarin Gal, and Dustin Tran. Uncertainty Baselines: Benchmarks for Uncertainty & Robustness in Deep Learning, 2022. URL http://arxiv.org/abs/2106.04015.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Brendan O'Donoghue, Botao Hao, Dieterich Lawson, Morteza Ibrahimi, Xiuyuan Lu, Vikranth Dwaracherla, and Benjamin Van Roy. The Neural Testbed: Evaluating Joint Predictions. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, volume 35 of *NIPS '22*, pp. 12554–12565, New Orleans, LA, USA, 2022. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/5141f6bc105d30edbae48f1d2e0b1e66-Paper-Conference.pdf.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, volume 32, Vancouver, Canada, 2019. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf.

Enrique Paredes and Rafael Ballester-Ripoll. SGEMM GPU Kernel Performance, 2017. URL https://doi.org/10.24432/C5MK70.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

Romain Pic, Clément Dombry, Philippe Naveau, and Maxime Taillardat. Proper Scoring Rules for Multivariate Probabilistic Forecasts Based on Aggregation and Transformation. *Advances in Statistical Climatology, Meteorology and Oceanography*, 11(1):23–58, 2025. URL https://ascmo.copernicus.org/articles/11/23/2025/.

Mingda Qiao and Eric Zhao. Truthfulness of Decision-Theoretic Calibration Measures, 2025. URL http://arxiv.org/abs/2503.02384.

R Quinlan. Auto MPG, 1993. URL https://doi.org/10.24432/C5859H.

Roberto Rossi, S. Armagan Tarim, Steven Prestwich, and Brahim Hnich. Piecewise linear approximations of the standard normal first order loss function. *Applied Mathematics and Computation*, 231:489–502, 2014. URL http://dx.doi.org/10.1016/j.amc.2014.01.019.

C. Sakar and Yomi Kastro. Online Shoppers Purchasing Intention Dataset, 2018. Published: UCI Machine Learning Repository.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520, Los Alamitos, CA, USA, June 2018. IEEE Computer Society. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00474.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2020. URL https://arxiv.org/abs/1910.01108.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized Affect Representations for Emotion Recognition. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3687–3697, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1404. URL https://aclanthology.org/D18-1404/.

Leonard J. Savage. Elicitation of Personal Probabilities and Expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1971.10482346. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1971.10482346.

Mark J. Schervish. A General Method for Comparing Probability Assessors. *The Annals of Statistics*, 17(4):1856–1879, 1989. URL https://www.jstor.org/stable/2241668.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the Pitfalls of Heteroscedastic Uncertainty Estimation with Probabilistic Neural Networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aPOpXlnV1T.

Novin Shahroudi, Viacheslav Komisarenko, and Meelis Kull. Aligning the Evaluation of Probabilistic Predictions with Downstream Value, 2025. URL http://arxiv.org/abs/2508.18251.

V Sigillito, S Wing, L Hutton, and K Baker. Ionosphere, 1989. URL https://doi.org/10.24432/C5W01B.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard (eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170/.

Gowthami Somepalli, Avi Schwarzschild, Micah Goldblum, C. Bayan Bruss, and Tom Goldstein. SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training. In *NeurIPS 2022 First Table Representation Workshop*, 2022. URL https://openreview.net/forum?id=FiyUTAy4sB8.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html.

Yan-Martin Tamm, Rinchin Damdinov, and Alexey Vasilev. Quality Metrics in Recommender Systems: Do We Calculate Metrics Consistently? In *Fifteenth ACM Conference on Recommender Systems*, pp. 708–713, 2021. URL https://doi.org/10.1145/3460231.3478848.

P Tfekci and H Kaya. Combined Cycle Power Plant, 2014. URL https://doi.org/10.24432/C5002N.

Athanasios Tsanas and Max Little. Parkinsons Telemonitoring, 2009. URL https://doi.org/10.24432/C5ZS3N.

Athanasios Tsanas and Angeliki Xifara. Energy Efficiency, 2012. URL https://doi.org/10.24432/C51307.

Meet P Vadera, Jinyang Li, Adam D Cobb, Brian Jalaian, Tarek Abdelzaher, and Benjamin M Marlin. URSABench: A System for Comprehensive Benchmarking of Bayesian Deep Neural Network Models and Inference methods. In *Proceedings of Machine Learning and Systems*, volume 4, pp. 217–237, Santa Clara, CA, USA, 2022. URL https://proceedings.mlsys.org/paper_files/paper/2022/file/4a420924d20bc025ebb37849169e6ebd-Paper.pdf.

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. URL https://journals.sagepub.com/doi/10.1177/0272989X06295361.

S Vito. Air Quality, 2008. URL https://doi.org/10.24432/C59K5F.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MINILM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. URL https://dl.acm.org/doi/abs/10.5555/3495724.3496209.

Marek Wydmuch, Kalina Jasinska, Mikhail Kuznetsov, Róbert Busa-Fekete, and Krzysztof Dembczyński. A No-Regret Generalization of Hierarchical Softmax to Extreme Multi-Label Classification. In *32nd Conference on Neural Information Processing Systems*, Montréal, Canada, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/8b8388180314a337c9aa3c5aa8e2f37a-Abstract.html.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms, 2017. URL https://arxiv.org/abs/1708.07747.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(41), 2023. URL https://doi.org/10.1038/s41597-022-01721-8.

Mochen Yang and Xuan Bi. Cost-Aware Calibration of Classifiers. *INFORMS Journal on Data Science*, 4(2):101–113, 2025. URL https://pubsonline.informs.org/doi/10.1287/ijds.2024.0038.

Ruotao Zhang, Constantine Gatsonis, and Jon Arni Steingrimsson. Role of calibration in uncertainty-based referral for deep learning. *Statistical Methods in Medical Research*, 32(5):927–943, 2023. URL https://journals.sagepub.com/doi/10.1177/09622802231158811.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level Convolutional Networks for Text Classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.

Kehao Zhu, Yingye Zheng, and Kwun Chuen Gary Chan. Weighted Brier Score—An Overall Summary Measure for Risk Prediction Models with Clinical Utility Consideration. *Statistics in Biosciences*, 2025. URL https://doi.org/10.1007/s12561-025-09505-5.

A Scope of the paper

This appendix clarifies the scope of our work along two axes: the *object* being evaluated, which we restrict to first-order probabilistic predictions, and the *decision-setting* in which it is used, which we restrict to one-shot and myopic sequential decisions.

First-order probabilistic predictions Throughout this paper, a probabilistic prediction refers to a single distribution over the label space; we call such an object a *first-order probabilistic prediction*. In probabilistic binary classification, this is a single $f \in [0, 1]$ encoding $\Pr(\mathrm{y} = 1)$; in regression, it is a single distribution over $\mathbb{Y}$ such as a Gaussian $\mathcal{N}(\mu, \sigma^2)$. This is the object almost universally produced and evaluated in probabilistic ML benchmarks (Ovadia et al., 2019; Nado et al., 2022; Basora et al., 2025), and the object for which all metrics we analyze are defined. A separate strand of the literature studies *second-order* probabilistic predictions: objects that take the form of a distribution over distributions. There are several reasons we restrict our scope to first-order evaluation. First, our work is positioned as a response to the current mainstream UQ evaluation literature (see sections 1 and 2), which consistently uses metrics defined on first-order probabilistic predictions. Second, there is no broadly accepted evaluation protocol for second-order UQ to begin with (Bengs et al., 2022; 2023). Many studies even evaluate the marginal predictive (see, e.g., Manchingal et al., 2025b;a), which is again first-order UQ and our analysis readily applies. Third, even if future research shifts more strongly toward second-order probabilistic predictions—a development we would welcome—we expect first-order evaluation to remain important for two reasons: *interpretability*—a single predictive distribution is substantially easier to communicate, visualize, and reason about than a distribution over distributions; and *decision-theoretic use*—Bayes-optimal actions under expected utility maximization are defined with respect to a single predictive distribution. Even models that maintain second-order beliefs internally are typically marginalized to a first-order posterior predictive distribution at decision time, so first-order quality remains the immediate driver of realized decision utility in most downstream pipelines.

Decision-setting The decision problems we consider are *one-shot*: an agent observes a prediction for an instance or batch, takes an action, and receives a utility that depends only on that action and the realized label. Sequential settings are covered insofar as they are *myopic*, i.e., decisions are made one instance/batch at a time, each maximizing the expected utility under the current predictive distribution, with no influence of the current action on future observations or future predictions. This is the setting of a large range of supervised deployment pipelines, e.g., screening, triage, or ranking, and it is the setting in which PSRs are characterized as expected utilities to begin with (Gneiting & Raftery, 2007); and it is also the setting of our electricity market case study.

Nevertheless, extending decision-aligned evaluation to second-order predictions and to non-myopic sequential decisions is an interesting and important direction which we leave for future work.

B Omitted related work

We first acknowledge work that shares similarities to our framework and goals and then summarize literature on the connection of calibration metrics and decision-making, to show why calibration metrics, which are important in their own right, do not fill the UQ evaluation gap PSRs leave. We close by stating what our method inherits from related work and what is explicitly novel.

B.1 Adjacent work

Ehm et al. (2016) study the evaluation and ranking of *point forecasts* through consistent scoring functions for functionals (e.g., quantiles). Similar to our decision-alignment, they consider *consistency* towards the functional of interest. They prove an order sensitivity result that resembles our proposition 3.2, however, they order point forecasts and associated scores, while we order scores and utilities of distributional reports. In addition, they show only one direction (consistency implies order sensitivity) and consider only inequalities, whereas we show that decision-alignment is *equivalent* to strict order and tie preservation. Further, Ehm et al. (2016) also discuss the decision-theoretic interpretation of binary decisions for PSRs and propose to use *Murphy diagrams* to evaluate and rank forecasts when the downstream task is unknown. As in our approach, they do not commit to a

single decision parameter θ. Murphy diagrams plot the score value over varying θ, whereas we set a plausible prior $\pi(\theta)$ and report the corresponding (scalar) integral value. In addition, we do not restrict ourselves to an inherent downstream task like binary decisions in binary classification, but offer several decision-specific metrics for several tasks.

Bradley & Schwartz (2011) also introduce a general framework tailored to ensemble weather forecasts, which is used to interpret known metrics. In their setup, probabilistic forecasts f for a continuous variable y are mapped to probability forecasts for threshold events of the form $\mathrm{y} \leq y$, and forecast quality is studied as a function $Q(y)$ of the threshold. Within this framework, several standard verification scores (e.g., the continuous ranked probability skill score) are shown to correspond to weighted averages of the forecast quality function over thresholds. Our decision-alignment framework is more general, allowing metric interpretation with respect to any given (integrable) utility family and is theoretically motivated through strict order and tie preservation in terms of expected downstream utility.

Shahroudi et al. (2025) propose to learn a proxy evaluation function for a known downstream task, motivated by the fact that evaluating the downstream utility can be costly. While their framework is termed *evaluation alignment*, similarly to our *decision-alignment*, the concepts differ substantially: In evaluation alignment, the decision loss needs to be known and is approximated with a learned PSR, while we propose general-purpose metrics for general-purpose evaluation where no specific downstream task is anticipated.

Ferrer & Ramos (2025) advocate the use of *expected PSRs (EPSRs)*, which are also defined as weighted integrals over a cost function (i.e., negative utility), but differ from our work in the sense that Ferrer & Ramos (2025) only consider classification with pointwise classification cost. They neither analyze other metrics in terms of such an integral representation nor propose concrete new weight functions (i.e., priors), nor consider other utilities. The focus of the paper is rather to demonstrate that such EPSRs are superior to calibration metrics for UQ evaluation.

Pic et al. (2025) propose another metric framework for probabilistic weather forecasts, specific to PSRs, based on aggregation (combining PSRs through a finite weighted sum) and transformation (transforming forecasts and labels before applying the PSR). This is a different framework than decision-alignment—we do not require a transformation and our aggregation step (prior-weighted integral) is more general.

Also, Derr & Williamson (2025) introduce a general framework for expressing evaluation metrics that represents metrics as the average realized value of gambles drawn from a predefined set. Their framework aims to unify a wide range of existing evaluation metrics by elucidating their structural relationships. In contrast, our work is not primarily concerned with unification, but with analyzing when and how metrics are informative for downstream decision-making.

Flores et al. (2025) also build on the Schervish representation (Schervish, 1989) to propose evaluation metrics for binary classifiers that reflect clinical priorities. Similarly, Flores et al. (2026) propose to use a clipped variant of the Brier score to evaluate binary classifiers to better reflect decision-relevant thresholds. These works are conceptually special cases of our PWU metrics.

B.2 Connection between calibration metrics and decision-making

Calibration metrics measure whether a model's predicted probabilities match observed frequencies, but this does not determine whether one model leads to better decisions than another. A constant predictor reporting the base rate is perfectly calibrated yet useless for any decision that requires discriminating between instances. Empirically, the link between calibration and decision quality is inconsistent. While some studies observed that good calibration implies good decisions (Laves et al., 2020; Zhang et al., 2023), others argue that metrics such as the ECE are flawed in terms of evaluating the usefulness of a model in downstream decisions (Gruber & Buettner, 2022; Kleinberg et al., 2023; Ferrer & Ramos, 2025). There has been substantial recent work on more decision-centric notions of calibration which aim to provide approximate guarantees (up to some tolerance ε) for downstream decision utilities or restricted families of tasks (Hu & Wu, 2024; Qiao & Zhao, 2025; Yang & Bi, 2025; Hegazy et al., 2025). These approaches are primarily motivated by training or post-hoc adjustment. While the calibration decision loss (CDL) of Hu & Wu (2024) can be used for model evaluation, it quantifies the cost of miscalibration, not the value of discrimination: the CDL is

zero for any calibrated model, including the uninformative base-rate predictor, and therefore cannot rank two calibrated models that differ in sharpness. In contrast, our PWU metrics are PSRs that reward both calibration and discrimination, and are designed to provide consistent model rankings in terms of downstream utility.

B.3 Novelty statement

Inherited from prior works The idea of the PWU construction is inspired by the integral representation of binary classification PSRs (as stated in section 4), and was also adopted by later works, e.g., Ferrer & Ramos (2025); Flores et al. (2025). Its reading as expected cost under an implicit weight over the cost ratio is established (Buja et al., 2005; Hernandez-Orallo et al., 2012). Furthermore, using a Beta-distributed cost-weighing has been proposed before, e.g., in Buja et al. (2005); Zhu et al. (2025). For regression, weighted integral representations also exist, e.g., for the continuous ranked probability score, but this relation is less general (Gneiting & Ranjan, 2011; Ehm et al., 2016).

Novel Our contribution is (1) the *decision-alignment* criterion as a joint property of (metric, utility family, prior), stated for arbitrary, possibly non-separable, dataset-level utilities, together with the separability barrier (lemma 3.5) and the uniqueness results for implicit priors (appendix G); (2) a systematic audit of 32 metric–utility pairs under this criterion, covering calibration and ranking metrics (ECE, MCE, R-AUC, E-Det) that have no integral representation, and utility families beyond thresholded binary classification, namely top-k selection, selective prediction with abstention cost, one-vs-rest multiclass decisions, and multivariate selective prediction; (3) concrete prior choices for these families, with the elicitation guidance and sensitivity analysis of appendices I.1 and J.3; and (4) an empirical study of ranking alignment and realized regret across 40 datasets, three modalities, and three applied case studies.

C Omitted proofs

C.1 Proof of Proposition 3.2

Proposition 3.2 (Decision-alignment $\Leftrightarrow$ strict order- and tie-preservation)**.** *Fix a decision family $\{U_\theta\}_{\theta\in\Theta}$ and a prior π. Then, for any fixed $\boldsymbol{y}$, the following statements are equivalent:*

(I) M is decision-aligned w.r.t. $\{U_\theta\}_{\theta\in\Theta}$ under π.

(II) M is strictly order-preserving and tie-preserving for $\{U_\theta\}_{\theta\in\Theta}$ under π, i.e., for all $\boldsymbol{f}_1, \boldsymbol{f}_2$,

$$M_{\boldsymbol{y}}(\boldsymbol{f}_1) < M_{\boldsymbol{y}}(\boldsymbol{f}_2) \Rightarrow \mathit{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}_1) < \mathit{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}_2),$$
$$M_{\boldsymbol{y}}(\boldsymbol{f}_1) = M_{\boldsymbol{y}}(\boldsymbol{f}_2) \Rightarrow \mathit{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}_1) = \mathit{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}_2).$$

Proof. Fix any $\boldsymbol{y}$.

(I)$\Rightarrow$(II): Since (I) holds, there exists a strictly increasing $h_{\boldsymbol{y}}$ s.t. $h_{\boldsymbol{y}}(M_{\boldsymbol{y}}(\boldsymbol{f})) = \mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f})$ for all $\boldsymbol{f}$. Take any $\boldsymbol{f}_1, \boldsymbol{f}_2$. If $M_{\boldsymbol{y}}(\boldsymbol{f}_1) < M_{\boldsymbol{y}}(\boldsymbol{f}_2)$, then strict monotonicity of $h_{\boldsymbol{y}}$ implies $h_{\boldsymbol{y}}(M_{\boldsymbol{y}}(\boldsymbol{f}_1)) < h_{\boldsymbol{y}}(M_{\boldsymbol{y}}(\boldsymbol{f}_2))$, hence $\mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}_1) < \mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}_2)$. If $M_{\boldsymbol{y}}(\boldsymbol{f}_1) = M_{\boldsymbol{y}}(\boldsymbol{f}_2)$, then, clearly, $h_{\boldsymbol{y}}(M_{\boldsymbol{y}}(\boldsymbol{f}_1)) = h_{\boldsymbol{y}}(M_{\boldsymbol{y}}(\boldsymbol{f}_2))$, hence $\mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}_1) = \mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}_2)$. Altogether, (II) holds.

(II)$\Rightarrow$(I): Assume (II) holds. Define $h_{\boldsymbol{y}}$ on $\mathrm{Im}(M_{\boldsymbol{y}})$ by

$$h_{\boldsymbol{y}}(m) := \mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}) \quad \text{for any } \boldsymbol{f} \text{ with } M_{\boldsymbol{y}}(\boldsymbol{f}) = m.$$

This is well-defined: if $M_{\boldsymbol{y}}(\boldsymbol{f}) = M_{\boldsymbol{y}}(\boldsymbol{f}')$, then tie preservation yields $\mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}) = \mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}')$. Moreover, for any $\boldsymbol{f}$, by construction $h_{\boldsymbol{y}}(M_{\boldsymbol{y}}(\boldsymbol{f})) = \mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f})$. It remains to show that $h_{\boldsymbol{y}}$ is strictly increasing. Let $m_1 < m_2$ be in $\mathrm{Im}(M_{\boldsymbol{y}})$, and choose $\boldsymbol{f}_1, \boldsymbol{f}_2$ s.t. $M_{\boldsymbol{y}}(\boldsymbol{f}_1) = m_1$ and $M_{\boldsymbol{y}}(\boldsymbol{f}_2) = m_2$. Then, strict order preservation implies $\mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}_1) < \mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}_2)$, i.e., $h_{\boldsymbol{y}}(m_1) < h_{\boldsymbol{y}}(m_2)$. Hence, $h_{\boldsymbol{y}}$ is strictly increasing, completing the proof. □

C.2 Proof of Lemma 3.5

Lemma 3.5 (Separability barrier)**.** *If the utility U_θ is pointwise-separable for every $\theta \in \Theta$ and $\theta \mapsto U_\theta(\boldsymbol{f}, \boldsymbol{y})$ is integrable for all $(\boldsymbol{f}, \boldsymbol{y})$, then any metric M that is decision-aligned w.r.t. $\{U_\theta\}_{\theta\in\Theta}$ (for some prior π) must be coordinate-independent.*

Proof. Fix any $\boldsymbol{y}$ and let M be decision-aligned w.r.t. $\{U_\theta\}_{\theta\in\Theta}$ for some prior weight π. Denote the instance function of U_θ by u_θ. By the assumed integrability of $\theta \mapsto U_\theta(\boldsymbol{f}, \boldsymbol{y})$ for all $(\boldsymbol{f}, \boldsymbol{y})$ and the fact that the sum in equation (2) is finite, we may interchange summation in equation (2) and integration in equation (1) (by linearity of the integral):

$$h_{\boldsymbol{y}}\big(M_{\boldsymbol{y}}(\boldsymbol{f})\big) = \int_\Theta -\frac{1}{n}\sum_{i=1}^{n} u_\theta(f_i, y_i)\pi(\theta)\mathrm{d}\theta = \frac{1}{n}\sum_{i=1}^{n}\int_\Theta -u_\theta(f_i, y_i)\pi(\theta)\mathrm{d}\theta.$$

Define the instance function $s(f_i, y_i) := \int_\Theta -u_\theta(f_i, y_i)\pi(\theta)\mathrm{d}\theta$, then the above equation reads

$$h_{\boldsymbol{y}}\big(M_{\boldsymbol{y}}(\boldsymbol{f})\big) = \frac{1}{n}\sum_{i=1}^{n} s(f_i, y_i).$$

Define $S_{\boldsymbol{y}}(\boldsymbol{f}) := \frac{1}{n}\sum_{i=1}^{n} s(f_i, y_i)$, which is (by the form of its definition) pointwise-separable and hence also coordinate-independent, so, for $a, b \in \mathbb{P}$ and $\boldsymbol{r}, \boldsymbol{r}' \in \mathbb{P}^{n-1}$,

$$S_{\boldsymbol{y}}((a, \boldsymbol{r})) < S_{\boldsymbol{y}}((b, \boldsymbol{r})) \Leftrightarrow S_{\boldsymbol{y}}((a, \boldsymbol{r}')) < S_{\boldsymbol{y}}((b, \boldsymbol{r}')),$$

and because $h_{\boldsymbol{y}}\big(M_{\boldsymbol{y}}(\boldsymbol{f})\big) = S_{\boldsymbol{y}}(\boldsymbol{f})$,

$$h_{\boldsymbol{y}}\big(M_{\boldsymbol{y}}((a, \boldsymbol{r}))\big) < h_{\boldsymbol{y}}\big(M_{\boldsymbol{y}}((b, \boldsymbol{r}))\big) \Leftrightarrow h_{\boldsymbol{y}}\big(M_{\boldsymbol{y}}((a, \boldsymbol{r}'))\big) < h_{\boldsymbol{y}}\big(M_{\boldsymbol{y}}((b, \boldsymbol{r}'))\big).$$

Since $h_{\boldsymbol{y}}$ is strictly increasing, it preserves orderings, and the same coordinate-independence property must hold for $M_{\boldsymbol{y}}$. The same statements follow analogously for equality and strict inequality in the opposite direction. □

C.3 Proof of Proposition 3.8

Proposition 3.8. *The NLL, BS, Acc, ECE, MCE, R-AUC, and E-Det are not decision-aligned w.r.t. $\{U_k\}_{k\in[n-1]}$.*

Proof. For any metric $M \in \{\text{NLL}, \text{BS}, \text{Acc}, \text{ECE}, \text{MCE}, \text{R-AUC}, \text{E-Det}\}$, assume for contradiction that there exists a prior $\pi : [n-1] \to \mathbb{R}_{\geq 0}$ s.t. for any fixed labels $\boldsymbol{y}$, there exists a strictly increasing function $h_{\boldsymbol{y}} : \mathrm{Im}(M_{\boldsymbol{y}}) \to \mathbb{R}$ s.t.

$$h_{\boldsymbol{y}}\big(M_{\boldsymbol{y}}(\boldsymbol{f})\big) = \sum_{k=1}^{n-1} -U_{k,\boldsymbol{y}}(\boldsymbol{f})\pi(k) \quad \forall \boldsymbol{f}. \tag{5}$$

The value of U_k is only influenced by the subset $\mathbb{S}_k(\boldsymbol{f})$, and we seek predictions $\boldsymbol{f}_1, \boldsymbol{f}_2$ that change our (monotonically transformed) metrics but not U_k. For this, consider the label vector $\boldsymbol{y} = \boldsymbol{0} = (0, \ldots, 0)$, because then, for every $k \in [n-1]$ and every $\boldsymbol{f}$, $U_{k,\boldsymbol{0}}(\boldsymbol{f}) = 0$, and equation (5) would imply that

$$h_{\boldsymbol{0}}\big(M_{\boldsymbol{0}}(\boldsymbol{f})\big) = 0 \quad \text{for all } \boldsymbol{f}. \tag{6}$$

Because $h_{\boldsymbol{0}}$ is strictly increasing, it is injective, and therefore equation (6) implies that $M_{\boldsymbol{0}}(\boldsymbol{f})$ is *constant* in $\boldsymbol{f}$: for all $\boldsymbol{f}_1, \boldsymbol{f}_2$,

$$M_{\boldsymbol{0}}(\boldsymbol{f}_1) = M_{\boldsymbol{0}}(\boldsymbol{f}_2). \tag{7}$$

We now exhibit $\boldsymbol{f}_1, \boldsymbol{f}_2$ s.t. equation (7) fails. Some of the vectors below have repeated entries, which is admissible under the tie-breaking convention introduced with $\mathbb{S}_k(\boldsymbol{f})$; the argument does not depend on which rule is chosen, since $\boldsymbol{y} = \boldsymbol{0}$ makes U_k vanish for every selection.

Case $M \in \{\mathbf{NLL}, \mathbf{BS}, \mathbf{Acc}\}$: Let $\boldsymbol{f}_1 = (\frac{1}{4}, \ldots, \frac{1}{4})$ and $\boldsymbol{f}_2 = (\frac{3}{4}, \ldots, \frac{3}{4})$. Remember that we consider the negative conventional accuracy as Acc. Then,

$$\begin{aligned} \mathrm{NLL}_{\mathbf{0}}(\boldsymbol{f}_1) = -\log(\tfrac{3}{4}) &\neq -\log(\tfrac{1}{4}) = \mathrm{NLL}_{\mathbf{0}}(\boldsymbol{f}_2), \\ \mathrm{BS}_{\mathbf{0}}(\boldsymbol{f}_1) = \tfrac{1}{16} &\neq \tfrac{9}{16} = \mathrm{BS}_{\mathbf{0}}(\boldsymbol{f}_2), \\ \mathrm{Acc}_{\mathbf{0}}(\boldsymbol{f}_1) = -1 &\neq 0 = \mathrm{Acc}_{\mathbf{0}}(\boldsymbol{f}_2), \end{aligned}$$

so equation (7) is violated.

Case $M \in \{\mathbf{ECE}, \mathbf{MCE}\}$: Fix any binning scheme $\mathcal{B} = \{B_1, \ldots, B_m\}$ that partitions $[0, 1]$ and let $\boldsymbol{f}_1, \boldsymbol{f}_2$ be as above. Then all entries of $\boldsymbol{f}_j$ fall into a single bin B_j containing the constant value. Since $\boldsymbol{y} = \mathbf{0}$, we have $\mathrm{freq}(B_j) = 0$ and $\mathrm{conf}(B_j)$ equals the constant value. Therefore,

$$\mathrm{ECE}_{\mathbf{0}}(\boldsymbol{f}_1) = |0 - \tfrac{1}{4}| = \tfrac{1}{4} \quad \neq \quad \tfrac{3}{4} = \mathrm{ECE}_{\mathbf{0}}(\boldsymbol{f}_2),$$

contradicting equation (7). Moreover, since only one bin is nonempty, the same holds for the MCE.

Case $M =$ R-AUC: Let $\boldsymbol{f}_1, \boldsymbol{f}_2$ be as above. We have $\mathrm{Acc}_{\mathbf{0}}(\boldsymbol{f}_{1,i:i\in R}) = -1$ for every retained set R, hence $\text{R-AUC}_{\mathbf{0}}(\boldsymbol{f}_1) = 0$. For $\boldsymbol{f}_2$, we have $\mathrm{Acc}_{\mathbf{0}}(\boldsymbol{f}_{2,i:i\in R}) = 0$ for every R, so the error rate is 1 on every retained set and $\text{R-AUC}_{\mathbf{0}}(\boldsymbol{f}_2) = \frac{1}{n}\sum_{r=0}^{n-1} \frac{n-r}{n} = \frac{n+1}{2n} > 0$, contradicting equation (7).

Case $M =$ E-Det: Note that for the E-Det with $\boldsymbol{y} = 0$, we have $e_i = 1$ iff $f_i > 0.5$. Choose numbers $a, b \in (0, 0.5)$ with $a < b$, and define $a' := 1 - a$, $b' := 1 - b$ so that $a', b' \in (0.5, 1)$ and $s(a) = s(a')$, $s(b) = s(b')$ by symmetry of entropy. Define two prediction vectors:

$$\boldsymbol{f}_1 = (\,a,\, b',\, b', \ldots, b'\,), \quad \boldsymbol{f}_2 = (\,b,\, a',\, a', \ldots, a'\,).$$

Then, for $\boldsymbol{f}_1$, we have exactly one index with $f_i \leq 0.5$ (hence $e_i = 0$) and $n - 1$ indices with $f_i > 0.5$ (hence $e_i = 1$). Moreover, the unique $e_i = 0$ point has uncertainty score $s(a)$, while every $e_i = 1$ point has uncertainty score $s(b') = s(b)$. Since $a < b$ and $s(\cdot)$ is strictly increasing on $(0, 0.5)$ for entropy, we have $s(a) < s(b)$, so all positives ($e = 1$) have strictly larger scores than all negatives ($e = 0$). Therefore, the AUROC of using s_i to predict e_i satisfies $\text{E-Det}_{\mathbf{0}}(\boldsymbol{f}_1) = -1$. For $\boldsymbol{f}_2$, we again have exactly one index with $e_i = 0$ and $n - 1$ indices with $e_i = 1$. Now the unique $e_i = 0$ point has uncertainty score $s(b)$, while every $e_i = 1$ point has uncertainty score $s(a') = s(a)$. Thus $s(b) > s(a)$ implies that the single negative has strictly larger score than all positives, so $\text{E-Det}_{\mathbf{0}}(\boldsymbol{f}_2) = 0$, contradicting equation (7).

In all cases, we reach a contradiction. Therefore, no such prior can exist, which proves the claim. □

D Relationship between decision-alignment and properness

In this section, we prove that many decision-aligned metrics (those where the transformations $h_{\boldsymbol{y}}$ are affine) are proper. Note, however, that the converse is not true. Indeed, **any PSR M is decision-aligned to *some* decision-family**, which we can simply define as $U_\theta = -\theta M$ on $\Theta = \mathbb{R}_{>0}$ (and considering $\pi(\theta) = \delta_1(\theta)$). Conversely, a PSR is not decision-aligned to *every* utility family. We discuss several such examples in this paper (for example, the NLL is not decision-aligned w.r.t. the top-k utility family).

Proposition D.1 (Decision-alignment with affine $h_{\boldsymbol{y}} \Rightarrow$ properness)**.** *Let the metric M be decision-aligned w.r.t. $\{U_\theta\}_{\theta\in\Theta}$ under π with transformations $h_{\boldsymbol{y}}$, let $\boldsymbol{f}^* \in \mathbb{P}^n$ denote the ground truth, and write $a_\theta(\boldsymbol{f})$ for a Bayes act under $\boldsymbol{f}$. Assume that*

(i) $\int_\Theta \mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}\left[|U_{\theta,\mathbf{y}}(\boldsymbol{f})|\right]\pi(\theta)\mathrm{d}\theta < \infty$ *for all $\boldsymbol{f} \in \mathbb{P}^n$, and*

(ii) there exist a constant $a > 0$ and a function $b : \mathbb{Y}^n \to \mathbb{R}$ with $\mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}\left[|b(\mathbf{y})|\right] < \infty$, s.t. $h_{\boldsymbol{y}}(m) = am + b(\boldsymbol{y})$ for all $\boldsymbol{y}$.

Then, M is a PSR. If, in addition, $\{U_\theta\}_{\theta\in\Theta}$ separates predictions *under π, that is, if*

$$\pi\left(\Theta_{\boldsymbol{f}^*,\boldsymbol{f}}\right) > 0 \quad \textit{for all } \boldsymbol{f} \in \mathbb{P}^n,\ \boldsymbol{f} \neq \boldsymbol{f}^*, \tag{8}$$

where $\Theta_{\boldsymbol{f}^*,\boldsymbol{f}} := \{\theta \in \Theta \mid a_\theta(\boldsymbol{f}^*)$ is the unique *Bayes act under $\boldsymbol{f}^*$ and $a_\theta(\boldsymbol{f}) \neq a_\theta(\boldsymbol{f}^*)\}$, then M is strictly proper.*

Proof. Define $S(\boldsymbol{f}, \boldsymbol{y}) := \mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f})$. By Dawid & Musio (2014), we know that $-U_{\theta,\boldsymbol{y}}(\boldsymbol{f})$ is a PSR for all θ. We first show that this implies that $S(\boldsymbol{f}, \boldsymbol{y})$ is a PSR as well. Denote by $\boldsymbol{f}^*$ the true distribution on $\mathbb{Y}^n$. By definition of properness, it holds for all predictions $\boldsymbol{f}$ that

$$\mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}\left[-U_{\theta,\mathbf{y}}(\boldsymbol{f}^*)\right] \le \mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}\left[-U_{\theta,\mathbf{y}}(\boldsymbol{f})\right] \quad \text{for all } \theta \in \Theta.$$

Multiplying by the nonnegative $\pi(\theta)$ and integrating both sides w.r.t. θ gives

$$\int_\Theta \mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}\left[-U_{\theta,\mathbf{y}}(\boldsymbol{f}^*)\right]\pi(\theta)\mathrm{d}\theta \le \int_\Theta \mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}\left[-U_{\theta,\mathbf{y}}(\boldsymbol{f})\right]\pi(\theta)\mathrm{d}\theta,$$

where the integrals exist by assumption (i). By Fubini-Tonelli (again using assumption (i)),

$$\int_\Theta \mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}\left[-U_{\theta,\mathbf{y}}(\boldsymbol{f})\right]\pi(\theta)\mathrm{d}\theta = \mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}\left[\int_\Theta -U_{\theta,\mathbf{y}}(\boldsymbol{f})\pi(\theta)\mathrm{d}\theta\right] = \mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}\left[S(\boldsymbol{f},\mathbf{y})\right].$$

With the same equality on $\boldsymbol{f}^*$ instead of $\boldsymbol{f}$, we get

$$\mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}\left[S(\boldsymbol{f}^*,\mathbf{y})\right] \le \mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}\left[S(\boldsymbol{f},\mathbf{y})\right],$$

which shows that $S(\boldsymbol{f}, \boldsymbol{y})$ is itself a PSR. By assumption (ii),

$$aM(\boldsymbol{f},\boldsymbol{y}) + b(\boldsymbol{y}) = \mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}) \;\Leftrightarrow\; M(\boldsymbol{f},\boldsymbol{y}) = \frac{1}{a}\mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}) - \frac{1}{a}b(\boldsymbol{y}),$$

we have for any prediction $\boldsymbol{f}$,

$$\mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}[M(\boldsymbol{f},\mathbf{y})] = \frac{1}{a}\mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}[S(\boldsymbol{f},\mathbf{y})] - \frac{1}{a}\mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}[b(\mathbf{y})].$$

so the set of minimizers of $\mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}[M(\boldsymbol{f},\mathbf{y})]$ coincides with that of $\mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}[S(\boldsymbol{f},\mathbf{y})]$. Thus, M is proper.

Now let $\boldsymbol{f} \neq \boldsymbol{f}^*$ and define the per-decision utility gap

$$g(\theta) := \mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}\left[-U_{\theta,\mathbf{y}}(\boldsymbol{f})\right] - \mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}\left[-U_{\theta,\mathbf{y}}(\boldsymbol{f}^*)\right].$$

It is π-integrable by assumption (i), and step 1 shows $g(\theta) \ge 0$ for every $\theta \in \Theta$. For $\theta \in \Theta_{\boldsymbol{f}^*,\boldsymbol{f}}$, the action $a_\theta(\boldsymbol{f})$ differs from the *unique* maximizer $a_\theta(\boldsymbol{f}^*)$, so that comparison is strict, $g(\theta) > 0$. A nonnegative function that is strictly positive on a set of positive measure has a strictly positive integral, so equation (8) yields

$$\int_\Theta g(\theta)\pi(\theta)\mathrm{d}\theta > 0 \quad\Longleftrightarrow\quad \mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}\left[S(\boldsymbol{f}^*,\mathbf{y})\right] < \mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}\left[S(\boldsymbol{f},\mathbf{y})\right].$$

By step 3, the strictly increasing affine link carries this over to M, giving $\mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}[M(\boldsymbol{f}^*,\mathbf{y})] < \mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}^*}[M(\boldsymbol{f},\mathbf{y})]$, i.e., M is strictly proper. □

E Counterexample for the converse relation of Lemma 3.5

Lemma 3.5 shows that a *pointwise-separable* utility family forces any decision-aligned metric to be coordinate-independent. In contrast, a pointwise-separable (hence coordinate-independent) metric can be decision-aligned w.r.t. a family in which *every* member is non-separable.

Let $\mathbb{Y} = \mathbb{R}$ and let $\mathbb{P}$ collect the distributions on $\mathbb{R}$ with finite second moment, writing $\mu_i := \mathbb{E}_{\mathrm{y}\sim f_i}[\mathrm{y}]$ for the predictive mean of f_i and $\boldsymbol{\mu} := (\mu_1, \ldots, \mu_n)$. Consider the MSE,

$$M(\boldsymbol{f},\boldsymbol{y}) := \frac{1}{n}\sum_{i=1}^n (\mu_i - y_i)^2,$$

which is pointwise-separable with instance function $s(f, y) := (\mu(f) - y)^2$, and therefore coordinate-independent.

Let $\Theta = \{-1, +1\}$ with prior $\pi(+1) = \pi(-1) = \frac{1}{2}$, fix $\rho \in (0, 1)$, and let the action space be $\mathbb{R}^n$, i.e. one point prediction per instance. Define the action–outcome utilities

$$u_\theta(\boldsymbol{a},\boldsymbol{y}) := -\frac{1}{n}(\boldsymbol{a}-\boldsymbol{y})^\top A_\theta(\boldsymbol{a}-\boldsymbol{y}) = -\frac{1}{n}\left[\sum_{i=1}^n (a_i - y_i)^2 + 2\theta\rho\,(a_1 - y_1)(a_2 - y_2)\right],$$

Table F.1: Summary of our decision-alignment analysis for classification *(left)* and regression *(right)*. If the metric is decision-aligned, we state the implicit prior π. Else, we write **not aligned**. Each stated prior is unique up to a positive constant (appendix G). All UQ metrics we consider are either **not decision-aligned** or encode largely **implausible prior beliefs**.

Metric	Binary decision	Top-k select.
NLL BS	$\pi(c) = \frac{1}{c(1-c)}$ $\pi(c) = 2$	not aligned
Acc	$\pi(c) = 2\delta_{0.5}(c)$	not aligned
ECE MCE	not aligned	not aligned
R-AUC E-Det	not aligned	not aligned

Metric	Multiclass one-versus-rest
NLL BS	$\pi(j, c) = \frac{1}{c}$ $\pi(j, c) = 2$
ECE	not aligned

Metric	Selective pred.	Top-k select.
NLL	$\pi(\lambda) = \varepsilon\lambda^{-2}$ for $\lambda, \sigma^2/v > \varepsilon > 0$	not aligned
MSE	$\pi(\lambda) = \delta_\infty(\lambda)$, not aligned for $\lambda < \infty$	not aligned
ECE MCE	not aligned	not aligned
R-AUC E-Det	not aligned	not aligned

Metric	Multivariate selective prediction
NLL MSE	not aligned not aligned
ES	not aligned

where $A_\theta := I_n + \theta\rho\,(\boldsymbol{e}_1\boldsymbol{e}_2^\top + \boldsymbol{e}_2\boldsymbol{e}_1^\top)$. Both A_{+1} and A_{-1} are positive definite because $\rho < 1$, so each u_θ is a strictly concave quadratic in $\boldsymbol{a}$. The two utilities differ only in the sign of their cross term. Decomposing $\boldsymbol{a} - \mathbf{y} = (\boldsymbol{a} - \boldsymbol{\mu}) - (\mathbf{y} - \boldsymbol{\mu})$ and using $\mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}}[\mathbf{y}] = \boldsymbol{\mu}$, the cross terms vanish and

$$\mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}}\left[(\boldsymbol{a} - \mathbf{y})^\top A_\theta(\boldsymbol{a} - \mathbf{y})\right] = (\boldsymbol{a} - \boldsymbol{\mu})^\top A_\theta(\boldsymbol{a} - \boldsymbol{\mu}) + \mathbb{E}_{\mathbf{y}\sim\boldsymbol{f}}\left[(\mathbf{y} - \boldsymbol{\mu})^\top A_\theta(\mathbf{y} - \boldsymbol{\mu})\right],$$

where the second term does not depend on $\boldsymbol{a}$ and is finite by assumption on $\mathbb{P}$. Thus, as A_θ is positive definite, the unique Bayes act is $a_\theta(\boldsymbol{f}) = \boldsymbol{\mu}$ for *both* θ, so that, writing $d_i := \mu_i - y_i$,

$$U_{\theta,\boldsymbol{y}}(\boldsymbol{f}) = u_\theta(\boldsymbol{\mu}, \boldsymbol{y}) = -\frac{1}{n}\left[\sum_{i=1}^{n} d_i^2 + 2\theta\rho\, d_1 d_2\right].$$

Now we can see that every member of the family is non-separable: Fix $n = 2$, $\rho = \frac{1}{2}$ and $\boldsymbol{y} = \mathbf{0}$, so that $d_i = \mu_i$, and let $a, b \in \mathbb{P}$ have predictive means 0 and 1. Taking a second coordinate $\boldsymbol{r}$ with predictive mean 0 gives $U_{+1,\mathbf{0}}((a, \boldsymbol{r})) = 0 > -\frac{1}{2} = U_{+1,\mathbf{0}}((b, \boldsymbol{r}))$, while taking $\boldsymbol{r}'$ with predictive mean -2 gives $U_{+1,\mathbf{0}}((a, \boldsymbol{r}')) = -2 < -\frac{3}{2} = U_{+1,\mathbf{0}}((b, \boldsymbol{r}'))$. Changing the first coordinate from a to b therefore moves the utility in opposite directions depending on the remaining coordinate, so U_{+1} is coordinate-dependent and hence non-separable. The same two values arise for U_{-1} upon replacing $\boldsymbol{r}'$ by a second coordinate of predictive mean $+2$.

However, the cross-terms cancel under π, so the MSE decision-aligned w.r.t. $\{U_\theta\}_{\theta\in\Theta}$ under π with $h_{\boldsymbol{y}} = \mathrm{id}$, even though both U_{+1} and U_{-1} are non-separable.

F Continuation of our decision-alignment analysis

We now extend our analysis from section 3.2 to univariate regression, multiclass classification and multivariate regression. Analogous results for additional metrics and decision families can be derived using the same template.

F.1 Univariate Regression

In this section, we consider probabilistic univariate regression models. We assume a Gaussian predictive distribution[11] and consider predictions of the form $f = (\mu, \sigma^2) \in \mathbb{P} = \mathbb{R} \times \mathbb{R}_{>0}$,

[11] We assume a Gaussian distribution for simplicity, but any distribution with a well-defined mean and variance is admissible.

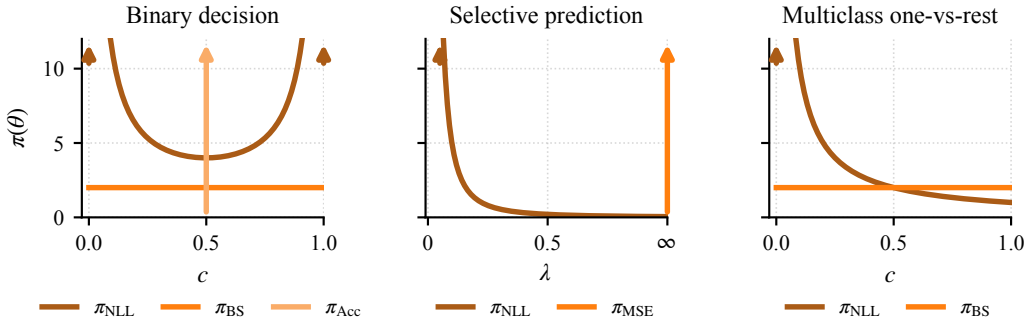


Figure F.1: Illustrations of the implausible priors identified in our decision-alignment analysis for the binary decision *(left)*, selective prediction *(center)*, and multiclass one-versus-rest decision *(right)* utility. The arrows indicate where the curves diverge.

consisting of a predictive mean $\mu \in \mathbb{R}$ and predictive variance $\sigma^2 \in \mathbb{R}_{>0}$. The regression target is the true label $y \in \mathbb{Y} = \mathbb{R}$. As evaluation metrics, we consider the NLL, mean squared error (MSE), ECE, MCE, R-AUC, and E-Det. We state the metric definitions in appendix K, where we also show that the **NLL and MSE are coordinate-independent**, and the **ECE, MCE, R-AUC, and E-Det are coordinate-dependent**. We summarize the results on the top right of table F.1.

F.1.1 Selective prediction

In selective prediction, the decision-maker chooses an action in $\{\text{predict}, \text{abstain}\}$, where selecting *predict* corresponds to issuing a point prediction $\hat{y}$, while *abstain* corresponds to deferring the decision (e.g., to a human expert). We assume squared loss for point predictions and a fixed abstention cost $\lambda \in \mathbb{R}_{>0}$. To put datasets of different scale on equal footing, we fix a reference scale $v > 0$ and express both the squared error and the predictive variance in units of v, writing

$$e := (\mu - y)^2 / v \quad \text{and} \quad r := \sigma^2 / v$$

for the *standardized* squared error and the *standardized* predictive variance, which renders the abstention cost $\lambda \in \mathbb{R}_{>0}$ dimensionless. Throughout, v is a fixed reference scale that does not depend on the outcomes being evaluated; in our experiments we use the training-target variance of the corresponding fold. With this setup, the Bayes act is to issue the predictive mean μ whenever the standardized predictive variance r is at most the abstention cost. This allows us to write the pointwise utility of a prediction $f = (\mu, \sigma^2)$ with true outcome $y \in \mathbb{R}$ for a fixed abstention cost λ as

$$u_\lambda(f, y) = -e\, \mathbf{1}_{r \le \lambda} - \lambda\, \mathbf{1}_{r > \lambda}.$$

Once again, the dataset-level utility $U_\lambda(\boldsymbol{f}, \boldsymbol{y})$ is defined as the average pointwise utility, so $\boldsymbol{U_\lambda}$ **is pointwise-separable**, and therefore the ECE, MCE, R-AUC, and E-Det are not decision-aligned w.r.t. $\{U_\lambda\}_{\lambda \in \mathbb{R}_{>0}}$ by lemma 3.5. We find that this also holds for the NLL and MSE.

Proposition F.1. *The NLL, MSE, ECE, MCE, R-AUC, and E-Det are not decision-aligned w.r.t.* $\{U_\lambda\}_{\lambda \in \mathbb{R}_{>0}}$.

Proof. For $M \in \{\text{ECE}, \text{MCE}, \text{R-AUC}, \text{E-Det}\}$, the claim follows directly by lemma 3.5. For $M \in \{\text{NLL}, \text{MSE}\}$, assume for contradiction that there exists a prior $\pi : \mathbb{R}_{>0} \to \mathbb{R}_{\ge 0}$ s.t. for any fixed labels $\boldsymbol{y}$ there exists a strictly increasing function $h_{\boldsymbol{y}} : \text{Im}(M_{\boldsymbol{y}}) \to \mathbb{R}$ s.t.

$$h_{\boldsymbol{y}}\big(M_{\boldsymbol{y}}(\boldsymbol{f})\big) = \int_{\mathbb{R}_{>0}} -U_{\lambda,\boldsymbol{y}}(\boldsymbol{f})\pi(\lambda)\mathrm{d}\lambda \quad \forall \boldsymbol{f}. \tag{9}$$

Case $M =$ MSE: Fix any labels $\boldsymbol{y}$ and consider two prediction vectors $\boldsymbol{f}_1 = (\boldsymbol{\mu}, \boldsymbol{\sigma_1^2})$, $\boldsymbol{f}_2 = (\boldsymbol{\mu}, \boldsymbol{\sigma_2^2})$ with identical predictive means $\boldsymbol{\mu} = \boldsymbol{y}$. Take any interval $[a, b) \subset \mathbb{R}_{>0}$ with $\pi([a, b)) > 0$ (must exist for a nonzero prior π) and consider standardized predictive variances $\sigma_{1,i}^2 / v = a < b = \sigma_{2,i}^2 / v$ for

all i. Then $\mathrm{MSE}(\boldsymbol{f}_1, \boldsymbol{y}) = \mathrm{MSE}(\boldsymbol{f}_2, \boldsymbol{y}) = 0$, since the MSE ignores predictive uncertainty entirely, so $h_{\boldsymbol{y}}\big(M_{\boldsymbol{y}}(\boldsymbol{f}_1)\big) = h_{\boldsymbol{y}}\big(M_{\boldsymbol{y}}(\boldsymbol{f}_2)\big)$. For the difference on the right-hand side of equation (9) when changing from $\boldsymbol{f}_1$ to $\boldsymbol{f}_2$, we have

$$\begin{aligned}&\int_{\mathbb{R}_{>0}} -U_{\lambda,\boldsymbol{y}}(\boldsymbol{f}_2)\pi(\lambda)\mathrm{d}\lambda - \int_{\mathbb{R}_{>0}} -U_{\lambda,\boldsymbol{y}}(\boldsymbol{f}_1)\pi(\lambda)\mathrm{d}\lambda \\ &= \int_{(0,b)} \lambda\,\pi(\mathrm{d}\lambda) + 0 - \int_{(0,a)} \lambda\,\pi(\mathrm{d}\lambda) - 0 = \int_{[a,b)} \lambda\,\pi(\mathrm{d}\lambda) \;\geq\; a\int_{[a,b)} \pi(\mathrm{d}\lambda) \;>\; 0,\end{aligned}$$

contradicting equation (9).

Case M = NLL: Fix any labels $\boldsymbol{y}$.

Step 1: Expected utility under π. Write $e_i := (\mu_i - y_i)^2/v$ and $r_i := \sigma_i^2/v$ for the standardized squared error and standardized predictive variance. For each coordinate i, $-U_{\lambda,y_i}(f_i) = e_i\mathbf{1}_{r_i\leq\lambda} + \lambda\mathbf{1}_{r_i>\lambda}$. By linearity of integration,

$$\int -U_{\lambda,\boldsymbol{y}}(\boldsymbol{f})\pi(\lambda)\mathrm{d}\lambda = \frac{1}{n}\sum_{i=1}^{n}\left(e_i\int_{r_i}^{\infty}\pi(\lambda)\mathrm{d}\lambda + \int_0^{r_i}\lambda\,\pi(d\lambda)\right). \tag{10}$$

Define $S(s) := \int_s^{\infty}\pi(\lambda)\mathrm{d}\lambda,\ s > 0$.

Step 2: One-dimensional reduction. Fix an index i. Fix all coordinates (μ_j, σ_j^2) for $j \neq i$ and fix $r_i = s \in (0,\infty)$. Let $x := e_i \in [0,\infty)$. Since the NLL depends on instance i only through $\frac{1}{2}\log\sigma_i^2 + \frac{e_i}{2r_i}$, and the reference scale cancels in the ratio e_i/r_i, we obtain under these choices

$$M_{\boldsymbol{y}}(\boldsymbol{f}) = A_s x + B_s, \quad A_s := \frac{1}{2ns} > 0, \tag{11}$$

where B_s is independent of x, and from equation (10),

$$\int -U_{\lambda,\boldsymbol{y}}(\boldsymbol{f})\pi(\lambda)\mathrm{d}\lambda = C_s x + D_s, \quad C_s := \frac{1}{n}S(s) \geq 0, \tag{12}$$

with D_s independent of x. Substituting (11)–(12) into equation (9) yields

$$h(A_s x + B_s) = C_s x + D_s \quad \forall x \geq 0. \tag{13}$$

Step 3: Form of h. Note the following: For $A > 0$, $B \in \mathbb{R}$, and h a function defined on $[B,\infty)$, we have that if there exist constants C, D s.t. $h(Ax + B) = Cx + D \quad \forall x \geq 0$, then

$$h(z) = \frac{C}{A}z + \left(D - \frac{C}{A}B\right) \quad \forall z \in [B,\infty).$$

This is because for any $z \geq B$, substituting $x = (z - B)/A \geq 0$ gives

$$h(z) = Cx + D = \frac{C}{A}z + \left(D - \frac{C}{A}B\right).$$

Applying this to equation (13), we conclude that

$$h(z) = m_s z + b_s \quad \forall z \geq B_s,$$

where

$$m_s := \frac{C_s}{A_s} = 2s\,S(s) \geq 0. \tag{14}$$

Because h is strictly increasing on $\mathrm{im}(M_{\boldsymbol{y}})$ and $[B_s,\infty) \subseteq \mathrm{im}(M_y)$, we must have $m_s > 0$, hence $S(s) > 0$ for all $s > 0$.

Step 4: Consistency across different variances. Let $s_1, s_2 > 0$. Repeating the above construction yields representations

$$h(z) = m_{s_1}z + b_{s_1} \quad (z \geq B_{s_1}), \quad h(z) = m_{s_2}z + b_{s_2} \quad (z \geq B_{s_2}).$$

The half-lines $[B_{s_1}, \infty)$ and $[B_{s_2}, \infty)$ overlap, and both expressions describe the same function h on the overlap. Therefore, $m_{s_1} = m_{s_2}$. Since s_1, s_2 are arbitrary, there exists a constant $m > 0$ s.t. $m_s = m \quad \forall s > 0$. Combining with equation (14) gives

$$S(s) = \frac{m}{2s} \quad \forall s > 0. \tag{15}$$

Step 5: Contradiction. From equation (15), we obtained

$$S(s) = \int_s^\infty \pi(\lambda)\mathrm{d}\lambda = \frac{m}{2s} \quad \forall s > 0,$$

for some constant $m > 0$. Since S is absolutely continuous with derivative $S'(s) = -\pi(s)$ for a.e. $s > 0$, differentiating yields

$$\pi(s) = -S'(s) = \frac{m}{2s^2} \quad \text{for a.e. } s > 0.$$

But then, for any fixed $s > 0$,

$$\int_0^s \lambda\pi(\lambda)\mathrm{d}\lambda = \int_0^s \lambda \cdot \frac{m}{2\lambda^2}\mathrm{d}\lambda = \frac{m}{2}\int_0^s \frac{1}{\lambda}\mathrm{d}\lambda = +\infty.$$

Consequently, the expected (negative) utility in equation (10) is infinite for any prediction vector with $\sigma_i^2 = s$ for some coordinate i, and thus the right-hand side of equation (9) is not well-defined, contradicting the definition of decision-alignment (which requires the integral to be well-defined for all admissible $\boldsymbol{f}$). Therefore, no such nonnegative measurable π can exist, and the regression NLL is not decision-aligned w.r.t. $\{U_\lambda\}_{\lambda>0}$. □

Notably, the proof of proposition F.1 reveals that under minor additional assumptions, the NLL *is* decision-aligned w.r.t. the U_λ-family under a Pareto prior.

Proposition F.2. *Fix any $\varepsilon > 0$ and define $\mathbb{P}_\varepsilon := \{f = (\mu, \sigma^2) \mid \sigma^2/v > \varepsilon\}$, i.e., predictions whose standardized predictive variance r exceeds ε. Consider only predictions $\boldsymbol{f} \in \mathbb{P}_\varepsilon^n$. Then, the NLL is decision-aligned w.r.t. $\{U_\lambda\}_{\lambda\in\mathbb{R}_{>\varepsilon}}$ under $\pi(\lambda) = \varepsilon\lambda^{-2}$ and*

$$h_{\boldsymbol{y}}(m) = 2\varepsilon m - \varepsilon\log(2\pi v\varepsilon) \quad \forall \boldsymbol{y}.$$

Proof. Write $e_i = (\mu_i - y_i)^2/v$ and $r_i = \sigma_i^2/v$ for the standardized squared error and standardized predictive variance. Restricting to $\lambda > \varepsilon$ and $f_i \in \mathbb{P}_\varepsilon$ for all i, the negative expected pointwise utility is

$$\int_\varepsilon^{r_i} \lambda\pi(\lambda)\mathrm{d}\lambda + e_i\int_{r_i}^\infty \pi(\lambda)\mathrm{d}\lambda.$$

Choose $\pi(\lambda)$ as the probability density function of a Pareto$(1, \varepsilon)$ distribution, given by $\pi(\lambda) = \varepsilon\lambda^{-2}$ for $\lambda \geq \varepsilon$. Substituting this density into the integral gives

$$\varepsilon\int_\varepsilon^{r_i}\frac{1}{\lambda}\mathrm{d}\lambda + \varepsilon\, e_i\int_{r_i}^\infty \frac{1}{\lambda^2}\mathrm{d}\lambda = \varepsilon\left(\log r_i - \log\varepsilon\right) + \varepsilon\frac{e_i}{r_i}.$$

Note that the reference scale cancels in the ratio, $e_i/r_i = (\mu_i - y_i)^2/\sigma_i^2$, while $\log r_i = \log\sigma_i^2 - \log v$. By linearity of expectation, the negative expected dataset-level utility is the average of the pointwise expected risks:

$$\begin{aligned}\int_\Theta -U_{\lambda,\boldsymbol{y}}(\boldsymbol{f})\pi(\lambda)\mathrm{d}\lambda &= \frac{1}{n}\sum_{i=1}^n\left[\varepsilon\log\sigma_i^2 + \varepsilon\frac{(\mu_i - y_i)^2}{\sigma_i^2} - \varepsilon\log v - \varepsilon\log\varepsilon\right] \\ &= 2\varepsilon\left[\frac{1}{n}\sum_{i=1}^n\left(\frac{1}{2}\log\sigma_i^2 + \frac{(\mu_i - y_i)^2}{2\sigma_i^2}\right)\right] - \varepsilon\log(v\varepsilon),\end{aligned} \tag{16}$$

which closely resembles the NLL. We can express equation (16) as an affine transformation of the NLL:

$$\int_\Theta -U_{\lambda,\boldsymbol{y}}(\boldsymbol{f})\pi(\lambda)\mathrm{d}\lambda = 2\varepsilon\cdot\mathrm{NLL}_{\boldsymbol{y}}(\boldsymbol{f}) - \varepsilon\log(2\pi v\varepsilon).$$

Defining $h_{\boldsymbol{y}}(x) = 2\varepsilon x - \varepsilon\log(2\pi v\varepsilon)$ for all $\boldsymbol{y}$, which is strictly increasing for $\varepsilon > 0$, we satisfy equation (1). Thus, $M = \mathrm{NLL}$ is decision-aligned with the selective prediction utility family under $\pi(\lambda) = \frac{\varepsilon}{\lambda^2}$ and $h_{\boldsymbol{y}}(m) = 2\varepsilon m - \varepsilon\log(2\pi v\varepsilon)$ for all $\boldsymbol{y}$. □

By proposition G.5 in appendix G, this Pareto$(1, \varepsilon)$ prior is unique up to a positive constant. Note that it is implausible, as it places more weight on λ the smaller λ is. In other words, it anticipates almost negligible abstention costs. We visualize the Pareto prior in the center of figure F.1, alongside the Dirac prior for the case $\lambda = \infty$ discussed below.

Remark F.3 (Case $\lambda = \infty$). When defining U_λ, we excluded the trivial case of $\lambda = \infty$ (always predict). Note, however, that the MSE equals $-U_\lambda$ for $\lambda = \infty$ up to the positive factor v (the standardization cancels under the strictly increasing $h_{\boldsymbol{y}}(m) = m/v$). In other words, the MSE is decision-aligned w.r.t. $\{U_\lambda\}_{\lambda \in \mathbb{R}_{>0} \cup \{\infty\}}$ under the (implausible) prior $\pi(\lambda) = \delta_\infty(\lambda)$, which is unique up to a positive constant (proposition G.5).

F.1.2 TOP-k SELECTION

We now consider the regression analog of top-k selection: pick k items to maximize return, with an additional penalty for outcomes falling below a reference level. As in selective prediction, the utility is evaluated on standardized outcomes: with the reference scale v of section F.1.1 and a reference location m, we write

$$\tilde{y} := \frac{y - m}{\sqrt{v}}, \qquad \tilde{\mu} := \frac{\mu - m}{\sqrt{v}}, \qquad \tilde{\sigma} := \frac{\sigma}{\sqrt{v}},$$

so that the reference level corresponds to $\tilde{y} = 0$ and the penalty applies to outcomes below it. In our experiments, m and $\sqrt{v}$ are the training-target mean and standard deviation of the corresponding fold, and hence do not depend on the outcomes being evaluated. Writing $(x)_+ := \max\{x, 0\}$, the utility is

$$U_\phi(\boldsymbol{f}, \boldsymbol{y}) = \frac{1}{k} \sum_{i \in \mathbb{S}_\phi(\boldsymbol{f})} \big[\tilde{y}_i - \gamma(-\tilde{y}_i)_+\big],$$

where $\phi = (k, \gamma) \in \Phi := [n-1] \times \mathbb{R}_{>0}$, and γ is a dimensionless penalty per unit of realized shortfall ($\gamma = 0$ is risk-neutral). The Bayes act selects the k instances with the largest expected item utility. For Gaussian reports $f_i = \mathcal{N}(\mu_i, \sigma_i^2)$, the expected shortfall is $\mathbb{E}[(-\tilde{\mathrm{y}}_i)_+] = -\tilde{\mu}_i \Phi_{\mathcal{N}}(z_i) + \tilde{\sigma}_i \varphi_{\mathcal{N}}(z_i)$ with $z_i := -\tilde{\mu}_i / \tilde{\sigma}_i$ (Rossi et al., 2014), so the selection score reads

$$q_i(\boldsymbol{f}) = \tilde{\mu}_i - \gamma\big[-\tilde{\mu}_i \Phi_{\mathcal{N}}(z_i) + \tilde{\sigma}_i \varphi_{\mathcal{N}}(z_i)\big],$$

where $\Phi_{\mathcal{N}}$ and $\varphi_{\mathcal{N}}$ denote the standard normal CDF and density. For $\gamma > 0$, a larger predictive variance lowers q_i, so the ranking trades off predicted benefit against downside risk.

As in classification, $\boldsymbol{U_\phi}$ **is coordinate-dependent**, so we cannot leverage lemma 3.5. Nevertheless, we can show that none of the metrics we consider is decision-aligned w.r.t. $\{U_\phi\}_{\phi \in \Phi}$.

Proposition F.4. *The NLL, MSE, ECE, MCE, R-AUC, and E-Det are not decision-aligned w.r.t.* $\{U_\phi\}_{\phi \in \Phi}$.

Proof. For any $M \in \{\text{NLL}, \text{MSE}, \text{ECE}, \text{MCE}, \text{R-AUC}, \text{E-Det}\}$, assume for contradiction that M is decision-aligned w.r.t. $\{U_\phi\}_{\phi \in \Phi}$, i.e., that there exists a prior π on Φ s.t. for any fixed labels $\boldsymbol{y}$ there exists a strictly increasing function $h_{\boldsymbol{y}} : \mathrm{Im}(M_{\boldsymbol{y}}) \to \mathbb{R}$ with

$$h_{\boldsymbol{y}}\big(M_{\boldsymbol{y}}(\boldsymbol{f})\big) = \int_\Phi -U_{\phi,\boldsymbol{y}}(\boldsymbol{f}) \pi(\phi)\, \mathrm{d}\phi \quad \forall \boldsymbol{f}.$$

Take the label vector $\boldsymbol{y} = \mathbf{0} = (0, \ldots, 0)$ and let $n \geq 2$. All labels coincide, so their standardized values coincide as well, $\tilde{y}_i = -m/\sqrt{v} =: c$ for all i. Whichever set $\mathbb{S}_\phi(\boldsymbol{f})$ is selected, every summand of U_ϕ therefore equals $c - \gamma(-c)_+$, so

$$U_{\phi,\mathbf{0}}(\boldsymbol{f}) = c - \gamma(-c)_+ =: \kappa(\phi) \qquad \text{for all } \boldsymbol{f} \in \mathbb{P}^n \text{ and all } \phi = (k, \gamma) \in \Phi.$$

On these labels the utility is thus independent of the prediction, and hence so is the expected utility under *any* prior π:

$$\int_\Phi -U_{\phi,\mathbf{0}}(\boldsymbol{f}) \pi(\phi) \mathrm{d}\phi = -\int_\Phi \kappa(\phi) \pi(\phi) \mathrm{d}\phi \qquad \text{for all } \boldsymbol{f} \in \mathbb{P}^n,$$

which is a constant independent of $\boldsymbol{f}$ (finite, since decision-alignment presupposes an integrable expected utility). Since $h_{\mathbf{0}}$ is strictly increasing, hence injective, decision-alignment of M would force

$$M_{\mathbf{0}}(\boldsymbol{f}) \text{ to be constant in } \boldsymbol{f}. \tag{17}$$

We now exhibit, for each metric, two predictions $\boldsymbol{f}_1, \boldsymbol{f}_2$ on which equation (17) fails, writing $\boldsymbol{f} = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ throughout.

Cases $M \in \{\text{MSE}, \text{NLL}\}$: Fix any $w > 0$ and compare

$$\boldsymbol{f}_1 := \big((0,\dots,0),(w,\dots,w)\big), \qquad \boldsymbol{f}_2 := \big((1,\dots,1),(w,\dots,w)\big).$$

Since $\boldsymbol{y} = \mathbf{0}$, we get $\text{MSE}_{\mathbf{0}}(\boldsymbol{f}_1) = 0 \neq 1 = \text{MSE}_{\mathbf{0}}(\boldsymbol{f}_2)$, and for the Gaussian NLL,

$$\text{NLL}_{\mathbf{0}}(\boldsymbol{f}_1) = \tfrac{1}{2}\log(2\pi w) \quad \neq \quad \tfrac{1}{2}\log(2\pi w) + \frac{1}{2w} = \text{NLL}_{\mathbf{0}}(\boldsymbol{f}_2).$$

Cases $M \in \{\text{ECE}, \text{MCE}\}$: Keep $\boldsymbol{f}_1$ from above and set $\boldsymbol{f}_3 := \big((a,\dots,a),(w,\dots,w)\big)$, where $a > 0$ is large enough that every observation lies outside every prediction interval. Under $\boldsymbol{f}_1$ each observation coincides with its predictive mean, so it is covered at every level and $\hat{\alpha}_j(\boldsymbol{f}_1, \mathbf{0}) = 1$ for all j; under $\boldsymbol{f}_3$ no observation is covered, so $\hat{\alpha}_j(\boldsymbol{f}_3, \mathbf{0}) = 0$ for all j. Writing $\bar{\alpha} := \frac{1}{m}\sum_{j=1}^{m} \alpha_j$, this gives

$$\text{ECE}_{\mathbf{0}}(\boldsymbol{f}_1) = 1 - \bar{\alpha}, \quad \text{ECE}_{\mathbf{0}}(\boldsymbol{f}_3) = \bar{\alpha}, \qquad \text{MCE}_{\mathbf{0}}(\boldsymbol{f}_1) = 1 - \min_j \alpha_j, \quad \text{MCE}_{\mathbf{0}}(\boldsymbol{f}_3) = \max_j \alpha_j,$$

so $\text{ECE}_{\mathbf{0}}(\boldsymbol{f}_1) \neq \text{ECE}_{\mathbf{0}}(\boldsymbol{f}_3)$ whenever $\bar{\alpha} \neq \frac{1}{2}$, and $\text{MCE}_{\mathbf{0}}(\boldsymbol{f}_1) \neq \text{MCE}_{\mathbf{0}}(\boldsymbol{f}_3)$ whenever $\min_j \alpha_j + \max_j \alpha_j \neq 1$. Both are conditions on the level family alone, and both hold for the levels $\alpha \in \{0.5, 0.8, 0.9, 0.95\}$ of appendix K, where $\bar{\alpha} = 0.7875$ and $\min_j \alpha_j + \max_j \alpha_j = 1.45$: the two ECE values are 0.2125 and 0.7875, and the two MCE values are 0.5 and 0.95. Should the levels be symmetric about $\frac{1}{2}$, so that both conditions fail, it suffices to replace $\boldsymbol{f}_3$ by a report covered at the wider levels but not at the narrower ones, which again separates both metrics.

Case $M =$ R-AUC: In regression, R-AUC is the retention-weighted average of the MSE on the retention sets R_r obtained by sorting instances by Gaussian predictive entropy. Comparing $\boldsymbol{f}_1$ and $\boldsymbol{f}_3$, the MSE equals 0 on every nonempty subset under $\boldsymbol{f}_1$ and a^2 under $\boldsymbol{f}_3$, so $\text{R-AUC}_{\mathbf{0}}(\boldsymbol{f}_1) = 0 \neq a^2\frac{n+1}{2n} = \text{R-AUC}_{\mathbf{0}}(\boldsymbol{f}_3)$, irrespective of how entropy ties are broken.

Case $M =$ E-Det: In regression, we use the Gaussian predictive entropy as an uncertainty score $s_i = s(\sigma_i^2)$ (strictly increasing in σ_i^2) and define error indicators by $e_i := \mathbf{1}_{\left|\frac{y_i - \mu_i}{|y_i| + \varepsilon}\right| > t}$ for a fixed relative-error threshold $t > 0$ and a fixed $\varepsilon > 0$. On $\boldsymbol{y} = \mathbf{0}$ this reduces to $e_i = \mathbf{1}_{|\mu_i| > t\varepsilon}$. Pick $0 < w_1 < w_n$, set $\boldsymbol{\sigma}^2 = (w_1, w_n, \dots, w_n)$ so that $s(w_1) < s(w_n)$, and define

$$\boldsymbol{f}_4 := \big((\,0,\ 2t\varepsilon, \dots, 2t\varepsilon\,), \boldsymbol{\sigma}^2\big), \qquad \boldsymbol{f}_5 := \big((\,2t\varepsilon,\ 0, \dots, 0\,), \boldsymbol{\sigma}^2\big).$$

As $2t\varepsilon > t\varepsilon$ strictly, both forecasts produce both error classes for $n \geq 2$. Under $\boldsymbol{f}_4$ we have $e_1 = 0$ and $e_i = 1$ for $i \geq 2$, so the unique point with $e = 0$ has score $s(w_1)$ while every point with $e = 1$ has score $s(w_n)$. Since $s(w_1) < s(w_n)$, all positives have strictly larger scores than the negative, hence $\text{E-Det}_{\mathbf{0}}(\boldsymbol{f}_4) = -1$. Under $\boldsymbol{f}_5$ we have $e_1 = 1$ and $e_i = 0$ for $i \geq 2$, so the unique positive has score $s(w_1)$ while all negatives have score $s(w_n)$, implying $\text{E-Det}_{\mathbf{0}}(\boldsymbol{f}_5) = 0$.

In all cases, we obtain a contradiction. Therefore no such prior π can exist, and M is not decision-aligned w.r.t. $\{U_\phi\}_{\phi \in \Phi}$. □

F.2 Multiclass classification

We consider probabilistic K-class classification predictions $f = (f^1, \dots, f^K) \in \mathbb{P} = \Delta^{K-1}$, $K > 2$, $\Delta^{K-1} := \{p \in [0,1]^K : \sum_{j=1}^{K} p^j = 1\}$, so f^j denotes the predicted probability that the true label $y \in \mathbb{Y} := [K] := \{1, \dots, K\}$ equals j.

As evaluation metrics, we consider the *multiclass* NLL, BS, and ECE. We define

$$\text{NLL}(\boldsymbol{f}, \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}\big(-\log f_i^{y_i}\big), \quad \text{BS}(\boldsymbol{f}, \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K}\big(f_i^j - \mathbf{1}_{y_i = j}\big)^2.$$

The multiclass ECE uses the standard extension to vector-valued predictions via one-versus-all binning (see, e.g., Guo et al., 2017). The NLL and BS are **pointwise-separable** (hence coordinate-independent) by definition. The multiclass ECE is **coordinate-dependent** by the same bin-coupling argument as in the binary case (Proposition K.1).

As a downstream task, we generalize the binary decision from section 3.2.1 to multiclass via a one-vs-rest decomposition. For each class $j \in [K]$, we consider the binary decision "flag instance as class j or not," with cost threshold $c \in (0,1)$. Defining the binary label $\tilde{y}^j := \mathbf{1}_{y=j}$, the Bayes act is to flag class j whenever $f^j > c$. The pointwise utility of a prediction f with true label y for class j at threshold c is

$$u_{j,c}(f,y) = -\tilde{y}^j(1-c)\mathbf{1}_{f^j \le c} - (1-\tilde{y}^j)c\mathbf{1}_{f^j > c}.$$

This is exactly the binary utility u_c from section 3.2.1, applied to the binary prediction–label pair $(f^j, \tilde{y}^j)$. The combined decision family is indexed by $\theta = (j,c) \in [K] \times (0,1)$, with dataset-level utility $U_{j,c}(\boldsymbol{f},\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^n u_{j,c}(f_i,y_i)$. In particular, $U_{j,c}$ is **pointwise-separable**.

Proposition F.5. *The multiclass NLL and BS are decision-aligned w.r.t.* $\{U_{j,c}\}_{j\in[K],c\in(0,1)}$ *with* $\pi_{\mathrm{NLL}}(j,c) = 1/c$ *and* $\pi_{\mathrm{BS}}(j,c) = 2$*, and* $h_{\boldsymbol{y}} = \mathrm{id}$ *for all* $\boldsymbol{y}$*.*

Proof. **NLL:** Evaluating the per-instance expected negative utility under $\pi(j,c) = 1/c$ gives

$$\begin{aligned}\sum_{j=1}^K \int_0^1 -u_{j,c}(f_i,y_i)\,\frac{1}{c}\,\mathrm{d}c &= \int_{f_i^{y_i}}^1 \frac{1-c}{c}\,\mathrm{d}c + \sum_{j\neq y_i}\int_0^{f_i^j} 1\,\mathrm{d}c \\ &= \left[-\log f_i^{y_i} - 1 + f_i^{y_i}\right] + \sum_{j\neq y_i} f_i^j \\ &= -\log f_i^{y_i} - 1 + f_i^{y_i} + (1 - f_i^{y_i}) = -\log f_i^{y_i}.\end{aligned}$$

In the last step we used the simplex constraint $\sum_{j\neq y_i} f_i^j = 1 - f_i^{y_i}$. Averaging over instances, $\mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}) = \mathrm{NLL}(\boldsymbol{f},\boldsymbol{y})$, so the NLL is decision-aligned with $h_{\boldsymbol{y}} = \mathrm{id}$.

BS: The multiclass BS decomposes as $\mathrm{BS}(\boldsymbol{f},\boldsymbol{y}) = \sum_{j=1}^K \mathrm{BS}^{(j)}(\boldsymbol{f},\boldsymbol{y})$, where $\mathrm{BS}^{(j)}(\boldsymbol{f},\boldsymbol{y})$ is the binary BS for the j-th one-vs-rest problem. By Proposition 3.6, each binary BS admits the integral representation (3) with prior $\pi(c) = 2$, giving $\mathrm{BS}(\boldsymbol{f},\boldsymbol{y}) = \sum_{j=1}^K \int_0^1 -U_{j,c}(\boldsymbol{f},\boldsymbol{y}) \cdot 2\ \mathrm{d}c$, which matches Definition 3.1 with $h_{\boldsymbol{y}} = \mathrm{id}$.

□

As in the binary case, the NLL and BS implicit priors are implausible (visualized on the right of figure F.1) and unique up to a positive constant (proposition G.6 in appendix G). The ECE is *not* decision-aligned as a direct consequence of lemma 3.5.

Corollary F.6. *The multiclass ECE is not decision-aligned w.r.t.* $\{U_{j,c}\}_{j\in[K],c\in(0,1)}$*.*

F.3 Multivariate regression

We consider probabilistic multivariate regression models with output dimension $D > 1$. We assume a Gaussian predictive distribution and consider predictions of the form $f = (\mu, \Sigma) \in \mathbb{P} = \mathbb{R}^D \times \mathbb{S}^D_{++}$, with predictive mean $\mu \in \mathbb{R}^D$ and predictive covariance $\Sigma \in \mathbb{S}^D_{++}$ (the cone of $D \times D$ symmetric positive definite matrices). The regression target is $y \in \mathbb{Y} := \mathbb{R}^D$.

As evaluation metrics, we consider the *multivariate* NLL, MSE, and the energy score (ES) (Gneiting & Raftery, 2007).[12] We define

$$\mathrm{NLL}(\boldsymbol{f},\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^n \left[\frac{D}{2}\log(2\pi) + \frac{1}{2}\log\det(\Sigma_i) + \frac{1}{2}(y_i-\mu_i)^\top \Sigma_i^{-1}(y_i-\mu_i)\right],$$

$$\mathrm{MSE}(\boldsymbol{f},\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^n \|y_i - \mu_i\|^2.$$

For a predictive distribution f and observation y, the per-instance ES is defined as

$$l_{\mathrm{es}}(f,y) = \mathbb{E}_{X\sim f}\|X - y\| - \tfrac{1}{2}\,\mathbb{E}_{X,X'\sim f}\|X - X'\|,$$

[12] To the best of our knowledge, there is no established notion of the ECE for multivariate regression. Therefore, we consider the ES, which is a more established metric in this context.

where X, X' are independent draws from f and $\|\cdot\|$ denotes the Euclidean norm, and the dataset-level ES is the average. All three metrics are **pointwise-separable** by definition.

As a downstream task, we generalize selective prediction from section F.1.1 to multivariate output by applying the univariate selective-prediction rule *coordinate-wise* to each standardized output dimension and averaging over dimensions. Concretely, for each output dimension $d \in [D]$ the decision-maker chooses an action in $\{\text{predict}, \text{abstain}\}$. We apply the same standardization as in the univariate case (dimension-wise), fixing a reference scale $v_d > 0$ per output dimension and writing

$$e_d := (y_d - \mu_d)^2 / v_d \quad \text{and} \quad r_d := \Sigma_{dd} / v_d$$

for the standardized squared error and the standardized marginal predictive variance in dimension d. The Bayes act in dimension d is to predict μ_d whenever the standardized predictive variance r_d is at most λ. The pointwise utility of a prediction $f = (\mu, \Sigma)$ with true outcome $y \in \mathbb{R}^D$ is

$$u_\lambda(f, y) = -\frac{1}{D} \sum_{d=1}^{D} \left[e_d \, \mathbf{1}_{r_d \le \lambda} + \lambda \, \mathbf{1}_{r_d > \lambda} \right].$$

The dataset-level utility $U_\lambda(\boldsymbol{f}, \boldsymbol{y})$ is defined as the average pointwise utility. In particular, U_λ is **pointwise-separable**.

Proposition F.7. *The multivariate NLL, MSE, and ES are not decision-aligned w.r.t.* $\{U_\lambda\}_{\lambda \in \mathbb{R}_{>0}}$.

Proof. Write $r_{i,d} := \Sigma_{i,dd}/v_d$ and $e_{i,d} := (y_{i,d} - \mu_{i,d})^2/v_d$ for the standardized marginal predictive variance and standardized squared error of instance i in dimension d. Consider any prior π. The per-instance expected negative utility takes the form

$$s(f_i, y_i) = \frac{1}{D} \sum_{d=1}^{D} \big[e_{i,d} \, S(r_{i,d}) + G(r_{i,d}) \big], \tag{18}$$

where $S(t) := \int_t^\infty \pi(\lambda)\, \mathrm{d}\lambda$ and $G(t) := \int_0^t \lambda \, \pi(\lambda)\, \mathrm{d}\lambda$. In particular, $s(f_i, y_i)$ depends on Σ_i only through its (standardized) diagonal $(r_{i,1}, \ldots, r_{i,D})$. We now show, for each metric, that a strictly monotone link to the expected utility cannot exist.

MSE: The MSE does not depend on Σ_i at all. Fix labels $\boldsymbol{y}$ and set $\mu_i = y_i$ for all i, so $\mathrm{MSE}(\boldsymbol{f}, \boldsymbol{y}) = 0$ regardless of the covariances. Then $e_{i,d} = 0$ for all i, d, and equation (18) reduces to $s(f_i, y_i) = \frac{1}{D} \sum_d G(r_{i,d})$. Since π is a nonzero measure on $\mathbb{R}_{>0}$, we can choose $0 < t_1 < t_2$ with $t_1, t_2 \neq 1$ and $\pi\big((t_1, t_2]\big) > 0$; then

$$G(t_2) - G(t_1) = \int_{(t_1, t_2]} \lambda \, \pi(\mathrm{d}\lambda) \;\ge\; t_1 \, \pi\big((t_1, t_2]\big) \;>\; 0.$$

Setting $\Sigma_i = t v_1 \, b_1 b_1^\top + \sum_{d \ge 2} v_d \, b_d b_d^\top$ (where b_d is the d-th standard basis vector), we obtain $r_{i,1} = t$ and $r_{i,d} = 1$ for $d \ge 2$. Choosing $t \in \{t_1, t_2\}$ yields two prediction vectors with identical MSE (both 0) but different expected utilities, contradicting the existence of a strictly increasing $h_{\boldsymbol{y}}$.

For the NLL and ES, we use the following shared construction. Since all three quantities (NLL, ES, and expected utility) are pointwise-separable, fixing $n = 2$, any labels $\boldsymbol{y}$, and f_2 arbitrarily, the strictly increasing function $h_{\boldsymbol{y}}$ would establish a strictly monotone link between the per-instance metric $l(f_1, y_1)$ and the per-instance expected utility $s(f_1, y_1)$. In particular, $l(f_1, y_1) \neq l(f_1', y_1)$ must imply $s(f_1, y_1) \neq s(f_1', y_1)$. Fix $\mu_1 = y_1$ and consider two covariance matrices that share the same diagonal but differ in their off-diagonal structure:

$$\Sigma_A = \mathrm{diag}(v_1, v_2, \ldots, v_D), \qquad \Sigma_B = \Sigma_A + \rho \sqrt{v_1 v_2} \, \big(b_1 b_2^\top + b_2 b_1^\top \big),$$

for some $\rho \in (0, 1)$ and $v_1 \neq v_2$ where b_d denotes the d-th standard basis vector. Both matrices are symmetric positive definite (the leading 2×2 block of Σ_B has determinant $v_1 v_2 (1 - \rho^2) > 0$). They share the same diagonal entries, hence the same standardized diagonal $(1, 1, \ldots, 1)$, and since $\mu_1 = y_1$ also $e_{1,d} = 0$ for all d. By equation (18), $s((y_1, \Sigma_A), y_1) = s((y_1, \Sigma_B), y_1) = \frac{1}{D} \sum_d G(1)$.

NLL: Since $\mu_1 = y_1$, the per-instance NLL reduces to $\frac{D}{2} \log(2\pi) + \frac{1}{2} \log\det(\Sigma_1)$. We have $\det(\Sigma_A) = \prod_d v_d$ and $\det(\Sigma_B) = v_1 v_2 (1 - \rho^2) \prod_{d \ge 3} v_d < \det(\Sigma_A)$, so the NLL takes different

values for Σ_A and Σ_B while the expected utility is the same. This contradicts the strictly monotone link.

ES: Since $\mu_1 = y_1$, the predictive distribution is $\mathcal{N}(y_1, \Sigma_1)$ and for $W := X - y_1$,

$$l_{\text{es}}\big((y_1, \Sigma_1), y_1\big) = (1 - 1/\sqrt{2})\, \mathbb{E}_{W\sim\mathcal{N}(0,\Sigma_1)}\|W\|.$$

Writing $\eta_1, \dots, \eta_D$ for the eigenvalues of Σ_1, by its eigendecomposition $\|W\|^2 \overset{d}{=} \sum_{d=1}^{D} \eta_d Z_d^2$ where $Z_1, \dots, Z_D$ are independent standard normal variables. We claim that the map $\Phi = (\eta_1, \dots, \eta_D) \mapsto \mathbb{E}\sqrt{\sum_{d=1}^{D} \eta_d Z_d^2}$ is strictly concave on $\mathbb{R}^D_{>0}$. For every fixed realization z, the map $\eta \mapsto \sqrt{\sum_d \eta_d z_d^2}$ is concave, as the composition of the concave, non-decreasing function $\sqrt{\cdot}$ with a linear function of η, so Φ is concave. Strict concavity of $\sqrt{\cdot}$ on $(0, \infty)$ then gives a strict inequality for almost every realization, and integrating preserves it: $\Phi(t\eta + (1-t)\eta') > t\,\Phi(\eta) + (1-t)\,\Phi(\eta')$. The eigenvalues of Σ_A are $(v_1, v_2, \dots, v_D)$, while those of Σ_B are $\big(\frac{v_1+v_2}{2} + \Delta,\ \frac{v_1+v_2}{2} - \Delta,\ v_3, \dots, v_D\big)$ with $\Delta := \sqrt{(\frac{v_1-v_2}{2})^2 + \rho^2 v_1 v_2} > |\frac{v_1-v_2}{2}|$. Hence Σ_B's eigenvalues are a strict mean-preserving spread of Σ_A's within the first two coordinates. By symmetry of Φ in its first two arguments, $\Phi(a) = \Phi(a')$ where $a' := (v_2, v_1, v_3, \dots, v_D)$, and similarly $\Phi(b) = \Phi(b')$ where b' swaps the first two entries of b. Both pairs share the same midpoint $\bar{a} := \frac{1}{2}(a + a') = \frac{1}{2}(b + b') = (\frac{v_1+v_2}{2}, \frac{v_1+v_2}{2}, v_3, \dots, v_D)$. Since $v_1 \neq v_2$ implies $a \neq a'$, and $\Delta > 0$ implies $b \neq b'$, strict concavity of Φ gives $\Phi(\bar{a}) > \Phi(a)$ and $\Phi(\bar{a}) > \Phi(b)$. Moreover, a and b lie on the line through $\bar{a}$ in the direction $(1, -1, 0, \dots, 0)$, with a at distance $|\frac{v_1-v_2}{2}|$ from $\bar{a}$ and b at distance $\Delta > |\frac{v_1-v_2}{2}|$, so $a = (1-t)\bar{a} + tb$ for $t := |\frac{v_1-v_2}{2}|/\Delta \in (0, 1)$ (up to swapping b and b', which by symmetry does not affect Φ). By strict concavity of Φ along this segment,

$$\Phi(a) > (1-t)\Phi(\bar{a}) + t\Phi(b) > (1-t)\Phi(b) + t\Phi(b) = \Phi(b),$$

where the second inequality uses $\Phi(\bar{a}) > \Phi(b)$. Hence $\mathbb{E}_{W\sim\mathcal{N}(0,\Sigma_A)}\|W\| = \Phi(a) > \Phi(b) = \mathbb{E}_{W\sim\mathcal{N}(0,\Sigma_B)}\|W\|$, so the ES takes different values for Σ_A and Σ_B while the expected utility is the same. This contradicts the strictly monotone link.

□

G Uniqueness of the identified priors

Whenever we established decision-alignment in this paper, we exhibited *one* prior π. We now show that this prior is essentially unique. Note first that definition 3.1 can never pin π down beyond a positive constant: if $(\pi, h_{\boldsymbol{y}})$ satisfies equation (1), then so does $(a\pi, a h_{\boldsymbol{y}})$ for every $a > 0$, and $a h_{\boldsymbol{y}}$ is still strictly increasing. We therefore say that a prior π is *unique* when it is *unique up to scaling*. Throughout, predictions are restricted to those for which the metric under consideration is finite.

G.1 Auxiliary lemmas

Our main tool is the classical fact that solutions of Cauchy's functional equation are linear under mild regularity conditions. We state the version we need on an interval.

Lemma G.1 (Cauchy's equation on an interval)**.** *Let $J \subseteq \mathbb{R}$ be an interval with $0 \in J$ and let $F : J \to \mathbb{R}$ satisfy $F(p+q) = F(p) + F(q)$ for all $p, q \in J$ with $p + q \in J$. If F is bounded on every compact subinterval of J, then there exists $a \in \mathbb{R}$ with $F(p) = ap$ for all $p \in J$.*

Lemma G.1 is classical: F satisfies Cauchy's equation on the interval $J/2$ (whose closure contains 0) and hence extends to an additive function on $\mathbb{R}$ (Kuczma & Gilányi, 2009, Theorem 13.5.3); an additive function on $\mathbb{R}$ that is bounded on some interval is continuous (Kuczma & Gilányi, 2009, Lemma 9.3.1 with (9.1.1)) and hence linear (Kuczma & Gilányi, 2009, equation (5.2.6)).

For a prior π and a family whose members are pointwise-separable with instance functions u_θ, we write

$$s_\pi(f, y) := \int_\Theta -u_\theta(f, y)\, \pi(\theta)\, \mathrm{d}\theta$$

for the instance-level NEU, so that $\mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f}) = \frac{1}{n}\sum_{i=1}^{n} s_\pi(f_i, y_i)$ whenever the instance terms are finite (by linearity of the integral, as in the proof of lemma 3.5). Since a metric need not be finite

on all of $\mathbb{P}$ for every label (the NLL, for instance, diverges on reports that assign probability 0 to the realized label), we write

$$\mathbb{P}_y := \{f \in \mathbb{P} \mid m(f, y) \text{ and } s_\pi(f, y) \text{ are finite}\}$$

for the reports *admissible at label* y, and restrict all statements below to these sets. In the one-versus-rest setting of proposition G.6, where $m(f, y) = -\log f^y$, this is $\mathbb{P}_y = \{f \in \Delta^{K-1} \mid f^y > 0\}$. The following lemma shows that, for separable pairs, the instance-level NEU of a decision-aligned metric is an *affine* function of the instance-level metric, whenever the metric can vary continuously.

Lemma G.2 (Affine link for separable pairs)**.** *Let M and every U_θ be pointwise-separable with instance functions m and u_θ, let M be decision-aligned w.r.t. $\{U_\theta\}_{\theta\in\Theta}$ under π, and assume that $\theta \mapsto u_\theta(f, y)$ is π-integrable for all (f, y).*

(a) Fix $y_0 \in \mathbb{Y}$ and let $I := \{m(f, y_0) : f \in \mathbb{P}_{y_0}\}$. If I is an interval or has at most two elements, then there exist $a > 0$ and $b \in \mathbb{R}$ s.t. $s_\pi(f, y_0) = a\, m(f, y_0) + b$ for all $f \in \mathbb{P}_{y_0}$.

(b) If $y_0 \neq y_1$ both satisfy the assumption of (a) with non-degenerate intervals I, then the constants a from (a) coincide for y_0 and y_1.

Proof. **(a)** Let $\boldsymbol{y} := (y_0, \dots, y_0)$ and $h := h_{\boldsymbol{y}}$. If $m(f, y_0) = m(f', y_0)$ for $f, f' \in \mathbb{P}_{y_0}$, then comparing $\boldsymbol{f} = (f, f_2, \dots, f_n)$ with $\boldsymbol{f}' = (f', f_2, \dots, f_n)$ in equation (1) gives $s_\pi(f, y_0) = s_\pi(f', y_0)$; similarly, $m(f, y_0) < m(f', y_0)$ implies $s_\pi(f, y_0) < s_\pi(f', y_0)$ because h is strictly increasing. Hence $G(m(f, y_0)) := s_\pi(f, y_0)$ defines a strictly increasing function $G : I \to \mathbb{R}$ on the admissible reports, and we must show that G is affine with positive slope. If $|I| \leq 2$, this is trivial. So let I be a non-degenerate interval. Fix $f_3, \dots, f_n \in \mathbb{P}$ and put $R := \sum_{i\geq 3} m(f_i, y_0)$ and $R' := \sum_{i\geq 3} s_\pi(f_i, y_0)$ (empty sums for $n = 2$). For $u, v \in I$ choose f_1, f_2 with $m(f_1, y_0) = u$ and $m(f_2, y_0) = v$; then equation (1) reads

$$H(u + v) = G(u) + G(v) \quad \forall u, v \in I, \qquad \text{where } H(t) := n\, h\big(\tfrac{t+R}{n}\big) - R'. \tag{19}$$

Fix $u_0 \in I$, let $J := I - u_0$ (an interval containing 0), and define $F(p) := G(p + u_0) - G(u_0)$ for $p \in J$. For $p, q \in J$ with $p + q \in J$, applying equation (19) once with $(u, v) = (p + u_0, q + u_0)$ and once with $(u, v) = (p + q + u_0, u_0)$ gives

$$F(p) + F(q) = H(p + q + 2u_0) - 2G(u_0) = F(p + q).$$

Since F is strictly increasing, lemma G.1 yields $F(p) = ap$ on J for some $a > 0$, i.e., $G(u) = au + b$ with $b := G(u_0) - au_0$.

(b) Let (a_0, b_0) and (a_1, b_1) be the constants from (a) for y_0 and y_1, and let I_{y_0}, I_{y_1} be the two images. Consider $\boldsymbol{y} := (y_0, y_1, y_0, \dots, y_0)$ with $f_3, \dots, f_n \in \mathbb{P}_{y_0}$ fixed and R, R' as above. For $u \in I_{y_0}$ and $v \in I_{y_1}$, choosing f_1, f_2 accordingly, equation (1) reads

$$n\, h_{\boldsymbol{y}}\big(\tfrac{u+v+R}{n}\big) = a_0 u + a_1 v + b_0 + b_1 + R'.$$

Taking interior points u, v and $\delta \neq 0$ small enough that $u + \delta \in I_{y_0}$ and $v - \delta \in I_{y_1}$ leaves the left-hand side unchanged, so $(a_0 - a_1)\delta = 0$. □

Finally, uniqueness questions can always be reduced to the NEU itself.

Lemma G.3 (Reduction to the NEU)**.** *Let M be decision-aligned w.r.t. $\{U_\theta\}_{\theta\in\Theta}$ under π with transformations $h_{\boldsymbol{y}}$. Then M is decision-aligned under a prior π' if and only if the metric $\mathrm{NEU}_{\theta\sim\pi} : (\boldsymbol{f}, \boldsymbol{y}) \mapsto \mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f})$ is decision-aligned under π'.*

Proof. $h_{\boldsymbol{y}}$ is a strictly increasing bijection from $\mathrm{Im}(M_{\boldsymbol{y}})$ onto $\mathrm{Im}(\mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}})$ with strictly increasing inverse. If $h'_{\boldsymbol{y}}(M_{\boldsymbol{y}}(\boldsymbol{f})) = \mathrm{NEU}_{\theta\sim\pi',\boldsymbol{y}}(\boldsymbol{f})$ for all $\boldsymbol{f}$, then $g_{\boldsymbol{y}} := h'_{\boldsymbol{y}} \circ h_{\boldsymbol{y}}^{-1}$ is strictly increasing with $g_{\boldsymbol{y}}(\mathrm{NEU}_{\theta\sim\pi,\boldsymbol{y}}(\boldsymbol{f})) = \mathrm{NEU}_{\theta\sim\pi',\boldsymbol{y}}(\boldsymbol{f})$. Conversely, if $g_{\boldsymbol{y}}$ is such a transformation for $\mathrm{NEU}_{\theta\sim\pi}$, then $g_{\boldsymbol{y}} \circ h_{\boldsymbol{y}}$ serves as $h'_{\boldsymbol{y}}$ for M. □

Hence, in all proofs below we may and do assume $M = \mathrm{NEU}_{\theta\sim\pi}$, which is pointwise-separable with instance function s_π whenever the family is. In particular, every uniqueness result below applies verbatim to the corresponding PWU metric.

G.2 BINARY DECISION

Proposition G.4 (Uniqueness for the binary decision family)**.** *Let π be a prior on $(0,1)$ that is either atomless or a single atom $\pi = w\delta_{c_0}$, and let M be decision-aligned w.r.t. $\{U_c\}_{c\in(0,1)}$ under π. Then π is unique up to scaling. In particular, the priors of the NLL, BS, and Acc in proposition 3.6 are unique up to scaling.*

Proof. Let M be decision-aligned under π'; by lemma G.3 we may assume $M = \mathrm{NEU}_{c\sim\pi}$. For label 0 we have $-u_c(f,0) = c\mathbf{1}_{f>c}$, hence

$$s_\pi(f,0) = \int_{(0,f)} c\,\pi(c)\,\mathrm{d}c =: G_\pi(f) \quad \text{and} \quad s_{\pi'}(f,0) = \int_{(0,f)} c\,\pi'(c)\,\mathrm{d}c =: G_{\pi'}(f)$$

for $f \in \mathbb{P} \supseteq (0,1)$, and both are finite on $(0,1)$ because the NEUs are. If π is atomless, G_π is continuous and nondecreasing on $(0,1)$, so the image I of $f \mapsto s_\pi(f,0)$ is an interval; if $\pi = w\delta_{c_0}$, then $G_\pi(f) = wc_0\mathbf{1}_{f>c_0}$ takes two values. By lemma G.2(a) with $y_0 = 0$, there are $a > 0$ and b with $G_{\pi'}(f) = aG_\pi(f) + b$ for all $f \in \mathbb{P}_0$. Letting $f \downarrow 0$ gives $b = 0$, since $G_\pi(f), G_{\pi'}(f) \to 0$. Thus the measures $c\,\pi'(c)\,\mathrm{d}c$ and $a\,c\,\pi(c)\,\mathrm{d}c$ agree on every interval $(0,f)$, $0 < f < 1$, so the two measures coincide. Dividing by $c > 0$ yields $\pi' = a\pi$. □

G.3 SELECTIVE PREDICTION

For selective prediction, uniqueness holds for *every* prior, including the restricted parameter spaces of proposition F.2 and the extended parameter space of remark F.3. We write $U_\infty(\boldsymbol{f}, \boldsymbol{y}) := -\frac{1}{n}\sum_i(\mu_i - y_i)^2/v$ for the "always predict" member (i.e., $\lambda = \infty$).

Proposition G.5 (Uniqueness for the selective prediction family)**.** *Let $\varepsilon \geq 0$, let Θ be either (ε,∞) or $(\varepsilon,\infty]$, consider predictions in $\mathbb{P}_\varepsilon = \{(\mu,\sigma^2) : \sigma^2/v > \varepsilon\}$, and let π be any prior on Θ. If M is decision-aligned w.r.t. $\{U_\lambda\}_{\lambda\in\Theta}$ under π, then π is unique up to scaling. In particular, the Pareto prior of the NLL (proposition F.2) and the prior δ_∞ of the MSE are unique up to scaling.*

Proof. Let M be decision-aligned under π'; by lemma G.3 we may assume $M = \mathrm{NEU}_{\lambda\sim\pi}$, and we write $g_{\boldsymbol{y}}$ for the corresponding strictly increasing transformations. Define the survival functions $S_\pi(s) := \pi(\Theta \cap [s,\infty])$ and $S_{\pi'}(s) := \pi'(\Theta \cap [s,\infty])$ for $s > \varepsilon$; they are finite because the NEUs are. Fix $\boldsymbol{y}$, an index i, all coordinates f_j, $j \neq i$, and a standardized predictive variance $\sigma_i^2/v = s > \varepsilon$, and let $f := (\mu_i - y_i)^2/v \in [0,\infty)$ denote the standardized squared error. Since $-u_\lambda(f_i, y_i) = f\mathbf{1}_{\lambda\geq s} + \lambda\mathbf{1}_{\lambda<s}$ (with $-u_\infty(f_i,y_i) = f$), we obtain

$$\mathrm{NEU}_{\lambda\sim\pi,\boldsymbol{y}}(\boldsymbol{f}) = \tfrac{1}{n}S_\pi(s)\,f + B_s \quad \text{and} \quad \mathrm{NEU}_{\lambda\sim\pi',\boldsymbol{y}}(\boldsymbol{f}) = \tfrac{1}{n}S_{\pi'}(s)\,f + B'_s,$$

where B_s, B'_s do not depend on f, and decision-alignment states $g_{\boldsymbol{y}}\big(\frac{1}{n}S_\pi(s)f + B_s\big) = \frac{1}{n}S_{\pi'}(s)f + B'_s$ for all $f \geq 0$. If $S_\pi(s) = 0$, the left-hand side does not depend on f, forcing $S_{\pi'}(s) = 0$. If $S_\pi(s) > 0$, then $g_{\boldsymbol{y}}$ is affine on the half-line $[B_s,\infty)$ with slope $S_{\pi'}(s)/S_\pi(s)$ (Step 3 in the proof of proposition F.1). Any two such half-lines overlap, and $g_{\boldsymbol{y}}$ is a single function, so this slope is a constant $a > 0$ that does not depend on s (nor on the fixed coordinates). Altogether, $S_{\pi'}(s) = aS_\pi(s)$ for all $s > \varepsilon$, where $a > 0$ is well-defined because $\pi \neq 0$ implies $S_\pi(s) > 0$ for some s. Consequently, $\pi' = a\pi$. □

The proof only uses that the instance-level NEU is affine in the standardized squared error with slope given by a survival function of π. This is equally true for the coordinate-wise multivariate family of appendix F.3 (fix (i,d) and vary $x = (y_{i,d} - \mu_{i,d})^2/v_d$, cf. equation (18)), so the prior of our multivariate selective prediction PWU metric is unique up to scaling as well.

G.4 MULTICLASS ONE-VERSUS-REST DECISION

Proposition G.6 (Uniqueness for the one-versus-rest family)**.** *Let $K \geq 3$, let π be a prior on $[K]\times(0,1)$ whose class-wise components $\pi(j,\cdot)$ are atomless measures on $(0,1)$, and let M be decision-aligned w.r.t. $\{U_{j,c}\}_{j\in[K],c\in(0,1)}$ under π. Then π is unique up to scaling. In particular, the priors $\pi_{NLL}(j,c) = 1/c$ and $\pi_{BS}(j,c) = 2$ of the multiclass NLL and BS are unique up to scaling.*

Proof. Let M be decision-aligned under π'; by lemma G.3 we may assume $M = \mathrm{NEU}_{(j,c)\sim\pi}$. For a prediction $f \in \Delta^{K-1}$ and a label y, the instance-level NEU reads

$$s_\pi(f, y) = A_y(f^y) + \sum_{j \neq y} G_j(f^j), \tag{20}$$

where $G_j(x) := \int_{(0,x)} c\, \pi(j, c)\, \mathrm{d}c$ and $A_j(x) := \int_{[x,1)} (1 - c)\, \pi(j, c)\, \mathrm{d}c$, and analogously for π' with G'_j, A'_j. Fix $y_0 \in [K]$. The set $\mathbb{P}$ is convex and contains $\{f \in \Delta^{K-1} : f^{y_0} > 0\}$ (it is Δ^{K-1} or this set for all metrics we consider); on it, all functions in equation (20) are finite because the NEUs are, and continuous since the $\pi(j, \cdot)$ are atomless. Hence the image of $f \mapsto s_\pi(f, y_0)$ is a non-degenerate interval (non-degenerate because $\pi \neq 0$), and lemma G.2(a) and (b) provide a constant $a > 0$ (independent of y_0) and constants b_{y_0} with

$$s_{\pi'}(f, y_0) = a\, s_\pi(f, y_0) + b_{y_0} \qquad \forall f \in \mathbb{P}_{y_0},\ \forall y_0 \in [K]. \tag{21}$$

Step 1: The class-wise components differ by a common $1/c$ term. Fix y_0 and two classes $j \neq l$ in $[K] \setminus \{y_0\}$ (possible as $K \geq 3$). For $u, w \geq 0$ with $u + w < 1$, apply equation (21) to $f = (1 - u - w)e_{y_0} + ue_j + we_l$ (e_i denote standard basis vectors) and set $D_j := G'_j - aG_j$, $D_l := G'_l - aG_l$ on $[0, 1)$. By equation (20),

$$D_j(u) + D_l(w) = \psi(u + w), \qquad \psi(z) := aA_{y_0}(1 - z) - A'_{y_0}(1 - z) + b_{y_0}.$$

Setting $w = 0$ (resp. $u = 0$) and using $D_l(0) = 0$ (resp. $D_j(0) = 0$) gives $D_j = \psi = D_l$ on $[0, 1)$, hence $\psi(u + w) = \psi(u) + \psi(w)$ whenever $u, w, u + w \in [0, 1)$. As $\psi = D_j$ is a difference of two finite nondecreasing functions, lemma G.1 yields $\psi(u) = \alpha_{y_0} u$ for some $\alpha_{y_0} \in \mathbb{R}$, i.e., $G'_j(u) = aG_j(u) + \alpha_{y_0} u$ for $u \in [0, 1)$ and all $j \neq y_0$. As in the proof of proposition G.4, this identifies the measures on $(0, 1)$:

$$\pi'(j, \cdot) = a\, \pi(j, \cdot) + \frac{\alpha_{y_0}}{c}\, \mathrm{d}c \qquad \forall j \neq y_0.$$

For two labels $y_0 \neq y_1$ and a class $j \notin \{y_0, y_1\}$ (again using $K \geq 3$), the two representations of $\pi'(j, \cdot)$ force $\alpha_{y_0} = \alpha_{y_1}$. Hence there is a single $\alpha \in \mathbb{R}$ with $\pi' = a\pi + \alpha\nu_0$, where $\nu_0(j, c) := 1/c$ for all j is the prior of the multiclass NLL.

Step 2: $\alpha = 0$. By the computation in the proof of proposition F.5, $s_{\nu_0}(f, y) = -\log f^y$. Inserting $\pi' = a\pi + \alpha\nu_0$ into equation (21) therefore gives $-\alpha \log f^{y_0} = b_{y_0}$ for all $f \in \mathbb{P}_{y_0}$, i.e., for all values $f^{y_0} \in (0, 1]$. This forces $\alpha = 0$ (and $b_{y_0} = 0$), so $\pi' = a\pi$. □

H PWU DOCUMENTATION

We now document our PWU metrics similarly to the model card framework (Mitchell et al., 2019).

Metric details PWU metrics are a family of PSRs for evaluating probabilistic predictions, introduced in section 4. A PWU metric is parameterized by a decision family $\{U_\theta\}_{\theta\in\Theta}$ and a prior π over Θ, and is defined as $\mathrm{PWU}(\boldsymbol{f}, \boldsymbol{y}) := \int_\Theta -U_\theta(\boldsymbol{f}, \boldsymbol{y})\, \pi(\theta)\, \mathrm{d}\theta$. The construction applies to any decision family with finite expected utility under π. We instantiate four PWU metrics in the binary and univariate setting: PWU-BD (binary decisions), PWU-SP (selective prediction), and two separate PWU-TS (top-k selection in classification and regression). We further introduce two analogous PWU instantiations for the multiclass and multivariate setting in sections F.2 and F.3.

Inputs and outputs A PWU metric takes a vector of first-order probabilistic predictions $\boldsymbol{f} \in \mathbb{P}^n$ together with the corresponding labels $\boldsymbol{y} \in \mathbb{Y}^n$, and returns a real-valued score (lower is better).

Intended use PWU metrics are intended for general-purpose uncertainty quantification benchmarking when no specific downstream task is given. They are designed as a complement to, not a replacement for, conventional UQ metrics: each PWU encodes a single decision family, so we recommend reporting a *set* of PWUs alongside standard scoring rules to cover a range of plausible downstream scenarios.

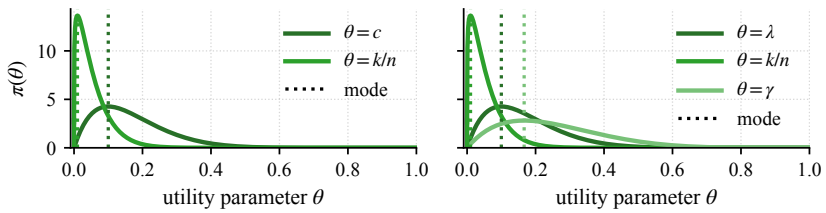


Figure I.1: Prior choices for our classification *(left)* and regression *(right)* PWU metrics.

Prior elicitation The prior π controls which regions of the decision space the metric emphasizes. We justify our specific prior choices and provide general elicitation guidance in appendix I; sensitivity to prior misspecification is analyzed in appendix J.3, where we find that PWU metrics retain robust positive utility alignment even under extreme prior perturbations.

Limitations No single PWU metric covers all downstream objectives; therefore, *several* PWU metrics are required to broadly represent a model's decision-making ability. Prior elicitation introduces a modeling step that generic metrics avoid. Our PWUs' priors may be misspecified in certain decision problems, which may degrade performance (even though we observed in appendix J.3 that PWU metrics exhibit robust positive alignment even under severe prior misspecification). The scope of PWU metrics is currently restricted to first-order probabilistic predictions (appendix A); extension to second-order UQ evaluation is left for future work. Computationally, evaluating a PWU metric requires integrating the negative utility against π, which requires numerical integration for prior choices that do not admit a closed-form. Mathematically, only utility–prior pairs are admitted that allow for a finite expected utility and hence a well-defined PWU.

Out-of-scope use PWU metrics are not intended as training objectives—some are non-separable, making mini-batching ill-defined, some are non-differentiable (top-k selection, thresholding), and the integral adds computational cost. They are also not intended as a fairness or robustness diagnostic: alignment with a chosen utility says nothing about performance on subpopulations not captured by that utility.

I Choosing PWU priors

Priors should be designed through consultation of the domain literature and expert knowledge. For each decision task, one should ask what range of decision parameters is plausible, and where the probability mass should lie. This process is neither straightforward nor unique—prior elicitation is a complex research area in its own right (see, e.g., Mikkola et al., 2024). Fortunately, PWU metrics are robust to misspecification; even PWUs corresponding to misspecified priors perform better than metrics corresponding to pathological priors, as demonstrated in our sensitivity analysis in appendix J.3.

In the remainder of this section, we provide concrete prior-elicitation guidance for benchmark designers who wish to construct their own PWU metrics. Appendix I.1 describes a four-step procedure for constructing a PWU metric in general, appendix I.2 lists default prior families, appendix I.3 provides a diagnostic prior-choice checklist, and appendix I.4 documents the prior choices used in our experiments as worked examples of the procedure.

I.1 A four-step PWU construction procedure

We recommend the following procedure for constructing a PWU prior.

Step 1: Identify the decision families Specify the action sets for potential downstream use and the corresponding utility (or loss) structure. Express this as parameterized families $\{U_\theta\}_{\theta \in \Theta}$.

Step 2: Identify the parameter spaces Θ Determine the units, range, and natural scale of θ. Is it bounded (cost ratios, selection fractions) or unbounded (abstention costs, risk-aversion coefficients)? For unbounded parameters, consider whether a natural rescaling exists (e.g., expressing abstention costs in units of a fixed reference scale v, as we do in section F.1.1). Choose such a reference scale independently of the labels being evaluated—we use the training-target variance of the corresponding fold—so that the utility does not depend on the outcomes it is scoring. Note that rescaling makes a parameter dimensionless but does not make its domain bounded: if the rescaled parameter still ranges over $\mathbb{R}_{>0}$, then choosing a prior supported on $(0, 1)$ is an additional modeling decision that restricts the decision family, and should be stated as such.

Step 3: Specify prior families Select parametric families for π_θ using the defaults in appendix I.2. For most decision parameters that arise in practice, a Beta distribution on a bounded or rescaled domain suffices.

Step 4: Set the prior's mode and concentration Pin down the mode from a single domain anchor (e.g., "a typical false negative is roughly $10\times$ as costly as a false positive" anchors c at ≈ 0.1). Set the concentration according to confidence in that anchor: a tight prior reflects strong belief in the modal value; a diffuse prior reflects broader uncertainty about the deployment context. We recommend verifying the choice using the diagnostics in appendix I.3.

I.2 Default prior families

Table I.1 summarizes default prior families for decision parameters that commonly arise in UQ benchmarking.

Table I.1: Default prior families for common decision parameter types. For unbounded scale parameters, we recommend rescaling to a bounded domain (e.g., dividing by the empirical variance of $\boldsymbol{y}$) and using a Beta prior, as we do for the abstention cost λ in section F.1.1.

Parameter type	Domain	Default family
Cost ratio / mixing weight	$(0, 1)$	Beta(α, β)
Selection fraction k/n	$(0, 1)$	Beta(α, β)
Abstention cost (rescaled)	$(0, 1)$	Beta(α, β)
Risk-aversion γ (dimensionless)	$(0, 1)$	Beta(α, β)
Unbounded scale parameter	$(0, \infty)$	LogNormal(μ, σ^2) or Gamma(α, β)

For a Beta(α, β) prior with $\alpha, \beta > 1$, the mode is $(\alpha - 1)/(\alpha + \beta - 2)$, and the concentration grows with $\alpha + \beta$.

I.3 Diagnostics

Before reporting PWU results, we recommend the following checks.

Visualization Plot the density of π_θ and verify that it concentrates mass where expected. We do this in figure I.1.

Quantile check Compute the 5th, 50th, and 95th percentiles of π_θ. Each should correspond to a plausible deployment scenario. For example, under our $\pi_c = \text{Beta}(2, 10)$, the 5th, 50th, and 95th percentiles of c are approximately 0.033, 0.148, and 0.364, corresponding to false-negative-to-false-positive cost ratios of roughly 29:1, 5.8:1, and 1.7:1. If any of these quantiles seems implausible for the target deployment context, the prior should be reconsidered.

Degeneracy check Verify that π_θ does not place substantial mass on regions where the decision becomes trivial or where UQ is least needed (e.g., $c \to 0$ or $c \to 1$ for binary decisions, $\lambda \to 0$ for selective prediction). This is exactly the pathology we identify in section 3 for the implicit priors of conventional metrics. If the prior is degenerate in this sense, truncate or otherwise reshape it.

Sensitivity check Re-run the benchmark with at least one perturbed prior (e.g., shifting the mode by 50%) and verify that conclusions are qualitatively stable. This makes the dependence of results on the prior transparent and guards against overfitting the prior to a particular outcome. Our sensitivity analysis in appendix J.3 provides a template for such checks.

I.4 Worked examples: prior choices used in our experiments

We document our specific prior choices below as worked examples of the procedure in appendix I.1. All priors were specified *a priori* from domain heuristics and held fixed across all experiments; they were not selected, tuned, or adjusted based on test-set performance, model rankings, or alignment results. The sensitivity analysis in appendix J.3 characterizes robustness to misspecification of these choices.

Classification For the binary decision problem, our PWU metric coincides with the well-known PSR integral representation equation (3) discussed in section 3.2.1. For the form (3), Buja et al. (2005) proposed PSRs that use Beta(α, β) priors over c. We follow this approach and set $\alpha = 2$ and $\beta = 10$. These are reasonable choices as the result is a distribution with modal value $c = 0.1$ and the majority of probability mass on small values of c. This reflects the cost asymmetry in many binary decisions where the cost of a false negative $1 - c$ is much higher than the cost of a false positive c. We use the same prior for the multiclass one-versus-rest decision. In the case of top-k selection, we place the prior over the selection *fraction* $k/n \in (0, 1)$ rather than over k itself, so that the same prior transfers across dataset sizes. We again propose a Beta distribution, this time targeting a lower mode of 0.01 (pick top 1%) as many top-k selection studies are about picking only a small number of items from a large dataset. Therefore, we set $\alpha = 1.2$ and $\beta = 20.8$, and write $\pi_{k/n} := \text{Beta}(1.2, 20.8)$ for the resulting density on $(0, 1)$.

Since the number of selected items is discrete while $\pi_{k/n}$ is a density, we now spell out how the prior weight $\pi_k(k)$ appearing in the definition of PWU-TS (section 4.1) is realized. We draw a fraction $q \sim \pi_{k/n}$ and round it up to the next whole item, $k = \lceil nq \rceil$, so that k receives the prior mass of all fractions that round up to it:

$$\pi_k(k) = F_{k/n}\left(\tfrac{k}{n}\right) - F_{k/n}\left(\tfrac{k-1}{n}\right), \qquad k = 1, \ldots, n-1,$$

where $F_{k/n}$ denotes the CDF of $\pi_{k/n}$. The residual mass above $(n-1)/n$ is assigned to $k = n - 1$, so that $k = 0$ and $k = n$ are excluded and the support matches the decision family $\{U_k\}_{k \in [n-1]}$.

Regression For selective prediction, we consider the abstention cost to be roughly 10% of the data variance ($\alpha = 2, \beta = 10$), so that abstention is often cheaper than the typical squared error but not negligible.[13] We use the same prior for the multivariate selective prediction PWU, applied dimension-wise with v_d the training-target variance of output dimension d. For top-k selection, we use the same prior as in classification for k/n, and say that the distribution for γ is parameterized by $\alpha = 2$ and $\beta = 6$. This corresponds to a moderate risk-aversion ($\gamma = 0$ corresponds to risk-neutrality). Since γ penalizes the realized shortfall of an already standardized outcome, it is dimensionless and is used as drawn, without any rescaling by the outcome variance. We write π_c, π_λ for Beta$(2, 10)$, π_k for Beta$(1.2, 20.8)$, and π_ϕ for the product density of Beta$(1.2, 20.8)$ and Beta$(2, 6)$ (assuming independence of k/n and γ).

On cross-validating the prior We do not cross-validate or otherwise tune the priors on labeled data. PWU priors are not hyperparameters in the conventional sense—they encode a belief about the downstream decision context, which is exogenous to the data. Selecting a prior to maximize alignment on a validation split would conflate "the prior best representing beliefs about a decision scenario" with "the prior under which a particular model looks best," reintroducing precisely the misalignment problem PWU metrics are designed to avoid. The appropriate analog of cross-validation is the sensitivity analysis in appendix J.3, which characterizes robustness to misspecification rather than tuning the prior.

[13] Standardizing by the reference scale v makes λ dimensionless, but it does not bound it: the natural domain of λ remains $\mathbb{R}_{>0}$. Placing a Beta prior on λ is therefore a deliberate restriction of the support to $(0, 1)$, expressing the belief that abstention costs less than the variance of the outcome. In practice, we draw λ on $(0, 1)$ and compare it against the standardized predictive variance σ^2/v, where v is the training-target variance of the corresponding fold—and hence independent of the labels being evaluated.

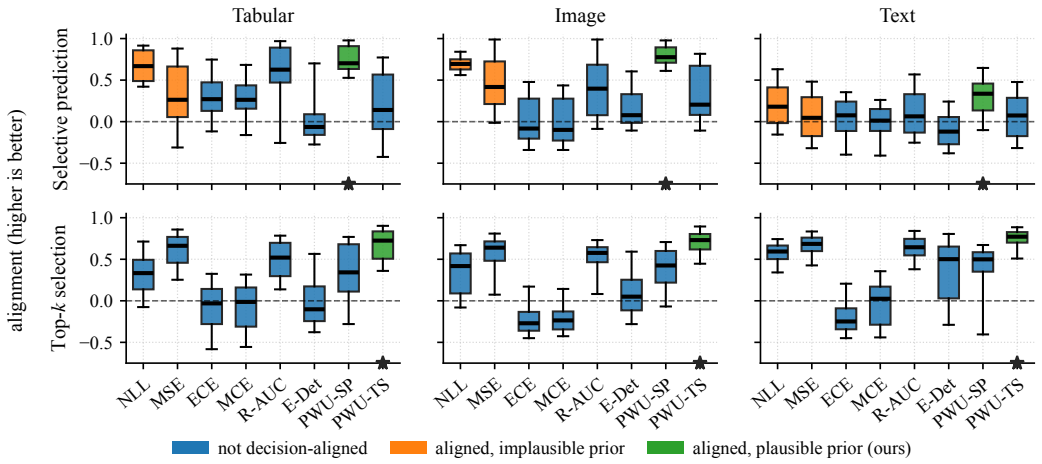


Figure J.1: Metric–utility alignment (Kendall's τ) in our univariate regression experiments. Each box's black line shows the median Kendall's τ value, the box spans the 25$^{\text{th}}$-75$^{\text{th}}$ percentile, and the whiskers the 5$^{\text{th}}$-95$^{\text{th}}$ percentile over 100 repeats; averaged over five datasets per modality. The coloring corresponds to our decision-alignment analysis. We mark the metric with the highest median with a star. Our PWU metrics align best with their corresponding utility family on every data modality.

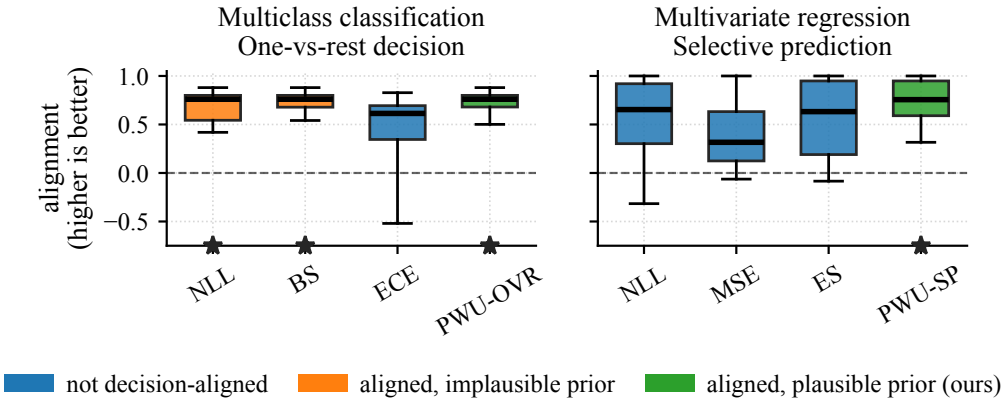


Figure J.2: Metric–utility alignment (Kendall's τ) in our multiclass and multivariate experiments. Each box's black line shows the median Kendall's τ value, the box spans the 25$^{\text{th}}$-75$^{\text{th}}$ percentile, and the whiskers the 5$^{\text{th}}$-95$^{\text{th}}$ percentile over 100 repeats; averaged over five datasets. The coloring corresponds to our decision-alignment analysis. We mark the metric with the highest median with a star. Our PWU metrics align (tied) best with their corresponding utility family.

J Details on our experiments

Our flagship experiments are binary classification and univariate regression experiments, which we run on tabular, image, and text data, and we provide an extension to the multiclass/multivariate setting on tabular data. We test six PWU metrics in total (one for each utility that appeared in this paper): PWU-BD (binary decision), PWU-TS (top-k selection—two variants: classification and regression), PWU-SP (selective prediction—two variants: univariate and multivariate), PWU-OVR (one-versus-rest decision). We report most results as figures. If desired, underlying numbers can be inspected in the folder `figure_numbers` in our codebase (provided as supplementary material).

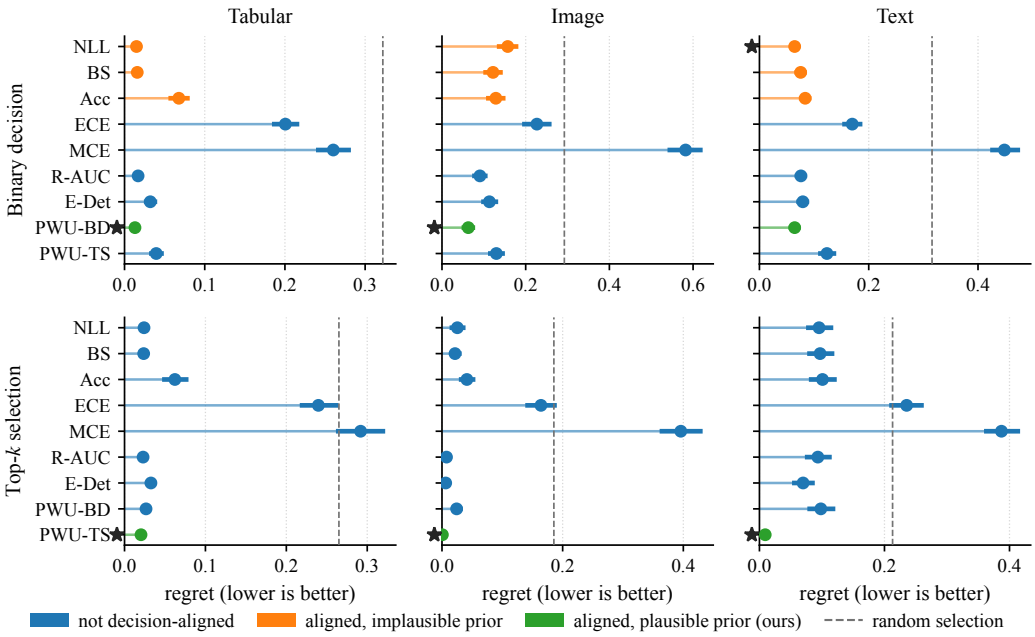


Figure J.3: Mean realized regret of model deployment as fraction of the achievable utility range (0 = always picks the best model), averaged over 100 repeats, in our binary classification experiments. The dot is the mean, and the bar a percentile bootstrap 95% confidence interval. The coloring corresponds to our decision-alignment analysis. We mark the metric with the lowest mean with a star. In 5/6 panels, one of our PWU metrics achieves lowest mean regret.

Model implementations For binary classification on tabular data, we implement logistic regression (LogReg), random forests (RFs), gradient boosting (GB), sparse variational Gaussian processes (GPs) (Hensman et al., 2015), multilayer perceptrons (MLPs), TabPFN (Hollmann et al., 2023), the FT-Transformer (Gorishniy et al., 2021), SAINT (Somepalli et al., 2022), the ResNet-MLP (Gorishniy et al., 2021), and a deep ensemble (Lakshminarayanan et al., 2017). For univariate regression on tabular data, we implement the same suite of models, replacing LogReg with linear regression (LinReg), GB with natural gradient boosting (NGB) (Duan et al., 2020), and MLPs with heteroscedastic MLPs. We then extend our experiments to multiclass/multivariate using the first five models from the respective suite.

For image classification, we implement logistic regression on raw pixels, a small CNN, ResNet-18 (He et al., 2016) trained from scratch—with and without data augmentation, and in variants with label smoothing (Szegedy et al., 2016), temperature scaling (Guo et al., 2017), and MC dropout (Gal & Ghahramani, 2016)—a deep ensemble of five ResNet-18s (Lakshminarayanan et al., 2017), and fine-tuned ImageNet-pretrained ResNet-18 and MobileNetV2 (Sandler et al., 2018).

For text classification, we implement TF-IDF features with LogReg and naive Bayes heads, frozen MiniLM sentence embeddings (Wang et al., 2020) with LogReg and MLP heads, fine-tuned DistilBERT (Sanh et al., 2020)—also in variants with temperature scaling, MC dropout, and a five-member deep ensemble—and fine-tuned BERT- (Devlin et al., 2019) and RoBERTa-Base (Liu et al., 2019).

For image and text regression, we mirror the two classification suites with Gaussian predictive distributions: every deep model outputs a mean and a standard deviation through a two-output head and is trained with the β-NLL (Seitzer et al., 2022), $\beta = 0.5$.

Our codebase is primarily based on the following packages: `scikit-learn` (Pedregosa et al., 2011) (BSD 3-Clause License), `gpytorch` (Gardner et al., 2018) (MIT license), `ngboost` (Duan et al., 2020) (Apache-2.0 license), and `tabpfn` (Hollmann et al., 2023) (Prior Labs License).

Further model details can be examined in our codebase, provided as supplementary material.

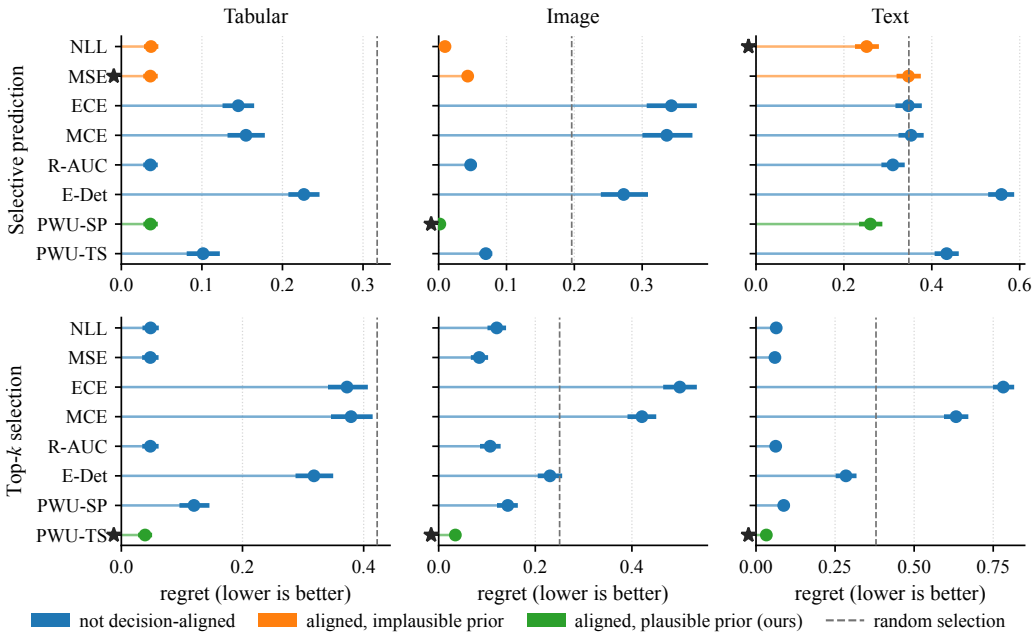


Figure J.4: Mean realized regret of model deployment as fraction of the achievable utility range (0 = always picks the best model), averaged over 100 repeats, in our univariate regression experiments. The dot is the mean, and the bar a percentile bootstrap 95% confidence interval. The coloring corresponds to our decision-alignment analysis. We mark the metric with the lowest mean with a star. In 4/6 panels, one of our PWU metrics achieves lowest mean regret.

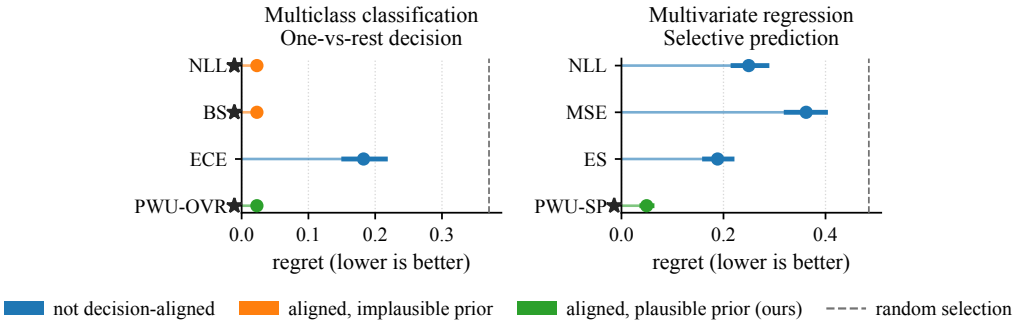


Figure J.5: Mean realized regret of model deployment as fraction of the achievable utility range (0 = always picks the best model), averaged over 100 repeats, in our multiclass and multivariate experiments. The dot is the mean, and the bar a percentile bootstrap 95% confidence interval. The coloring corresponds to our decision-alignment analysis. We mark the metric with the lowest mean with a star. In each panel, one of our PWU metrics achieves (tied) lowest mean regret.

Alignment score computation We now explain how we compute the alignment score for one metric–utility pair (M, U_θ) on one dataset, and this follows equally for all considered metric–utility pairs on each dataset. We train and test each model $\mathcal{M}_j$, $j = 1, \ldots, n_{\mathcal{M}}$ ($n_{\mathcal{M}} = 10$ in most of our experiments) on the dataset using 5-fold cross-validation (80/20 train–test partition) with fixed random seed, and repeat this train/test loop for 100 different seeds.[14] So for each model $\mathcal{M}_j$, we

[14]To keep the image and text suites (classification and regression) affordable, we avoid per-repeat refitting on these modalities: each of the 10 models per suite is trained once per seed (3 seeds) on the fixed training split of

obtain 100×5 metric/utility values: $\{m_{\text{rep,fold}}\}_{\mathcal{M}_j}, \{u_{\text{rep,fold}}\}_{\mathcal{M}_j}$, rep $= 1, \ldots, 100$, fold $= 1, \ldots, 5$. In each of the 100 repeats, we average the metric/utility value over the 5 folds to obtain a stable metric/utility value for each repeat: $\{\bar{m}_{\text{rep}}\}_{\mathcal{M}_j}, \{\bar{u}_{\text{rep}}\}_{\mathcal{M}_j}$. Using these 100 metric/utility values, we measure the *alignment* in each repeat by computing Kendall's τ between the two score vectors

$$\boldsymbol{s}_{M,\text{rep}} = \big(\bar{m}_{\text{rep},\mathcal{M}_1}, \ldots, \bar{m}_{\text{rep},\mathcal{M}_{n_\mathcal{M}}}\big), \qquad \boldsymbol{s}_{U_\theta,\text{rep}} = \big(\bar{u}_{\text{rep},\mathcal{M}_1}, \ldots, \bar{u}_{\text{rep},\mathcal{M}_{n_\mathcal{M}}}\big),$$

We compute τ from these values rather than from the induced rankings because several metric/utility pairs take the same value on different models, and ranking first would break such ties arbitrarily, in the order in which the models happen to be sorted. We therefore use the tie-corrected variant τ_b, which adjusts the denominator for the number of tied pairs. If a metric or a utility takes the same value on *all* models in a repeat, τ_b is undefined $(0/0)$; we score such cases as $\tau_b = 0$, reading them as exhibiting no association. Through this procedure, we obtain 100 metric-utility alignment scores $\{\tau_{(M,U_\theta),\text{rep}}\}$.

Regret computation The regret analysis reuses exactly the per-repeat values $\bar{m}_{\text{rep},\mathcal{M}_j}$ and $\bar{u}_{\text{rep},\mathcal{M}_j}$ of the previous paragraph. Within a repeat, a metric M selects the model it scores best,

$$\hat{\jmath}_{M,\text{rep}} := \arg\min_j \bar{m}_{\text{rep},\mathcal{M}_j}$$

(with $\arg\max$ for metrics where larger is better), and its *realized decision-regret* is the utility forgone by deploying that model instead of the utility-optimal one,

$$\text{reg}_{M,\text{rep}} := \max_j \bar{u}_{\text{rep},\mathcal{M}_j} \; - \; \bar{u}_{\text{rep},\mathcal{M}_{\hat{\jmath}_{M,\text{rep}}}}.$$

We report this relative to the utility spread of the model pool in that repeat,

$$\widetilde{\text{reg}}_{M,\text{rep}} := \frac{\text{reg}_{M,\text{rep}}}{\max_j \bar{u}_{\text{rep},\mathcal{M}_j} - \min_j \bar{u}_{\text{rep},\mathcal{M}_j}},$$

setting $\widetilde{\text{reg}}_{M,\text{rep}} := 0$ in the degenerate case where all models attain the same utility. The normalization is thus by the utility *range spanned by the candidate models*, not by the absolute utility of the best model, so that $\widetilde{\text{reg}}_{M,\text{rep}} \in [0, 1]$, taking the value 0 when the metric selects a utility-optimal model and 1 when it selects the worst one. The *Random* baseline replaces $\bar{u}_{\text{rep},\mathcal{M}_{\hat{\jmath}_{M,\text{rep}}}}$ by the pool mean $\frac{1}{n_\mathcal{M}} \sum_j \bar{u}_{\text{rep},\mathcal{M}_j}$, the expected utility of selecting a model uniformly at random. The value we report is the mean of that quantity, together with a percentile bootstrap 95% confidence interval over 2,000 resamples.

Choice of Kendall's τ Proposition 3.2 shows that decision-alignment is equivalent to strict order- and tie-preservation between the rankings induced by a metric and a utility. Therefore, evaluating decision-alignment reduces to assessing agreement between the two rankings. Standard measures for ranking similarity include Spearman's ρ and Kendall's τ. We choose Kendall's τ because it is robust and interpretable for small rankings, as in our case. For example, $\tau_{(M,U_\theta)} = 0.8$ means that 90% of all pairwise model comparisons induced by the metric M agree with the downstream utility U_θ. Other possibilities to compare two ranking vectors are top-k agreement or cardinal comparisons (e.g., Pearson's ρ). As discussed above, these do not evaluate decision-alignment in the sense of proposition 3.2. That said, these other metrics can be valid for different evaluation goals: top-k if the objective is identifying the best models, and cardinal correlations (e.g., Pearson's ρ) if magnitudes matter. So as an insightful ablation, we also report the top-1 and top-3 agreement and Pearson's ρ on a subset of our benchmark experiments (binary classification, univariate regression, tabular data) in figures J.6, J.7, and J.8. In each case, our PWU metrics are (tied) best.

J.1 Experiments on benchmark datasets

Datasets We use the following benchmark datasets from the UCI repository (Kelly et al., 2025): adult (Becker & Kohavi, 1996), bank marketing (Moro et al., 2014), ionosphere (Sigillito et al.,

each of the 5 datasets, and the 100 evaluation repeats are bootstrap subsamples of the held-out test split, so only $3 \times 5 = 15$ training runs per model are required, i.e., 150 per suite (somewhat more in practice, since the deep ensemble fits five constituent networks per run). Similarly, on the time series data in our electricity market case study, we use a different scheme for measuring variability, reported below.

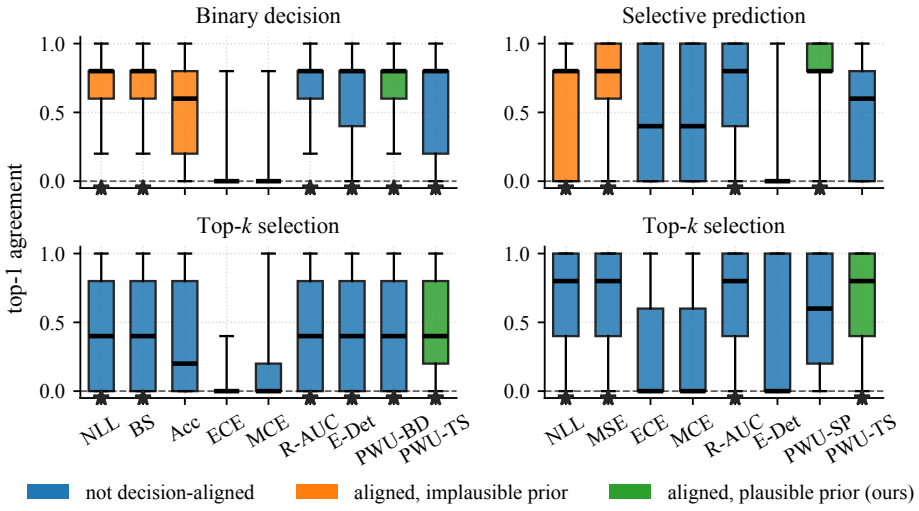


Figure J.6: Top-1 agreement in our binary classification *(left)* and univariate regression *(right)* experiments on tabular data. Each box's black line shows the median top-1 agreement value, the box spans the 25th-75th percentile, and the whiskers the 5th-95th percentile over 100 repeats; averaged over five datasets. The coloring corresponds to our decision-alignment analysis. Our PWU metrics show the (tied) strongest top-1 agreement.

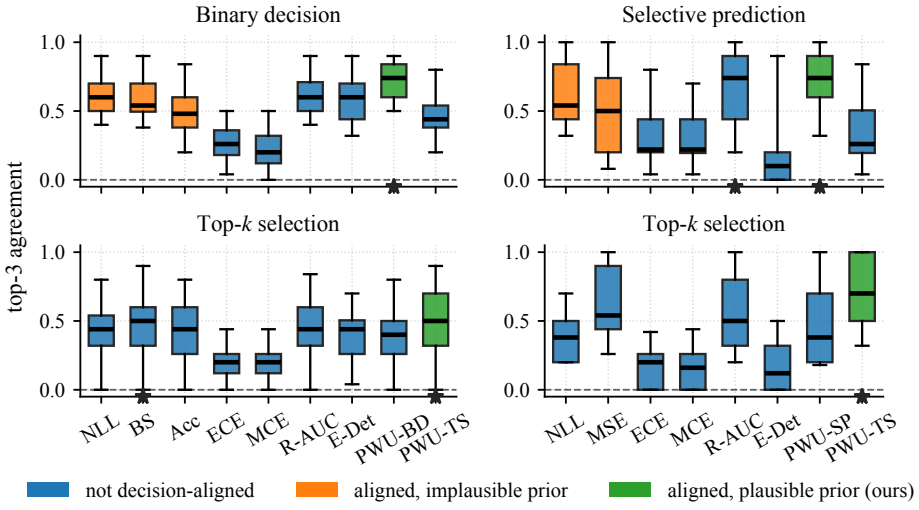


Figure J.7: Top-3 agreement in our binary classification *(left)* and univariate regression *(right)* experiments on tabular data. Each box's black line shows the median top-3 agreement value, the box spans the 25th-75th percentile, and the whiskers the 5th-95th percentile over 100 repeats; averaged over five datasets. The coloring corresponds to our decision-alignment analysis. Our PWU metrics show the (tied) strongest top-3 agreement.

1989), MAGIC gamma telescope (Bock, 2004), and online shoppers purchasing intention (Sakar & Kastro, 2018) for binary classification, and air quality (Vito, 2008), auto MPG (Quinlan, 1993), energy efficiency (Tsanas & Xifara, 2012), power plant (Tfekci & Kaya, 2014), and wine quality (Cortez et al., 2009) for univariate regression. For multiclass classification, we use the UCI datasets iris (Fisher, 1936), wine (Aeberhard & Forina, 1992), dry bean (Koklu & Özkan, 2020),

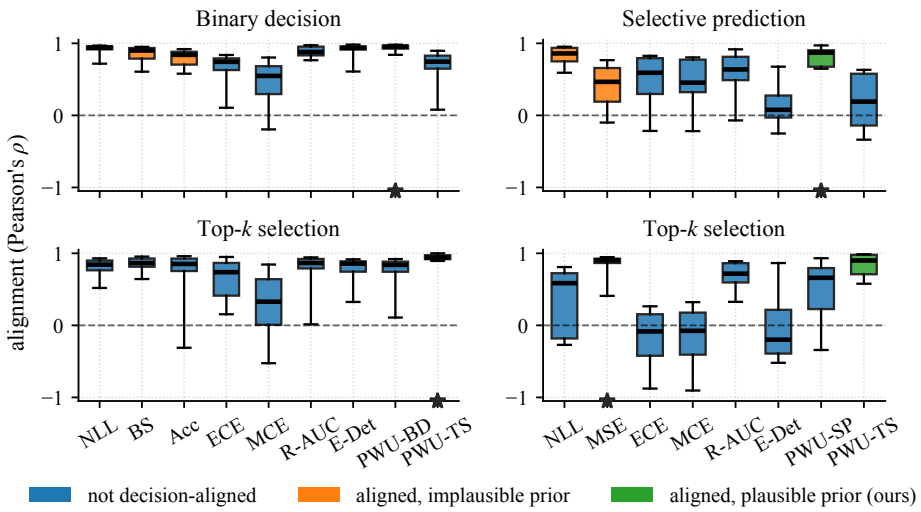


Figure J.8: Metric–utility correlation (Pearson's ρ) in our binary classification *(left)* and univariate regression *(right)* experiments on tabular data. Each box's black line shows the median Pearson's ρ value, the box spans the 25$^{\text{th}}$-75$^{\text{th}}$ percentile, and the whiskers the 5$^{\text{th}}$-95$^{\text{th}}$ percentile over 100 repeats; averaged over five datasets. The coloring corresponds to our theoretical findings from section 3. With one exception, our PWU metrics show the strongest ranking correlation.

pen-based recognition of handwritten digits (Alpaydin & Alimoglu, 1996), and covertype (Blackard, 1998); for multivariate regression, we use energy efficiency (Tsanas & Xifara, 2012), air quality (Vito, 2008), solar flare (Bradshaw, 1989), Parkinsons telemonitoring (Tsanas & Little, 2009), and SGEMM GPU kernel performance (Paredes & Ballester-Ripoll, 2017). For image classification, we use CIFAR-10 (Krizhevsky & others, 2009), Fashion-MNIST (Xiao et al., 2017), and the PneumoniaMNIST, DermaMNIST, and BloodMNIST datasets from the MedMNIST v2 collection (Yang et al., 2023). For text classification, we use IMDb (Maas et al., 2011), AG News (Zhang et al., 2015), CivilComments (Borkan et al., 2019), Emotion (Saravia et al., 2018), and the sentiment task of TweetEval (Barbieri et al., 2020). For image regression, we use the Global Wheat Head Detection dataset (David et al., 2020), RetinaMNIST from MedMNIST v2 (Yang et al., 2023), dSprites (Matthey et al., 2017), and rotated versions of Fashion-MNIST and CIFAR-10, predicting the rotation angle. For text regression, we use the continuous sentiment scores of the Stanford Sentiment Treebank (Socher et al., 2013), the query well-formedness ratings of Faruqui & Das (2018), the valence ratings of EmoBank (Buechel & Hahn, 2017), the toxicity fractions of CivilComments (Borkan et al., 2019), and the 1–10 star ratings of the IMDb reviews (Maas et al., 2011).

Utilities We measure the alignment of all metrics and utilities considered in section 3 and appendix F. We do not settle on one specific value for θ, but sample 5 θ_j-values from the corresponding priors π_θ, that we define in appendix I. We then average the alignment scores of a metric M with U_{θ_j}, $\tau_{(M,U_{\theta_j})}$ across the respective decision family to obtain $\tau_{(M,U_\theta)} = \frac{1}{5}\sum_{j=1}^{5} \tau_{(M,U_{\theta_j})}$.

We visualize the alignment results, averaged over all datasets, in figure 3 (binary classification), figure J.1 (univariate regression), and figure J.2 (multiclass and multivariate). The corresponding realized regret results can be found in figure J.3 (binary classification), figure J.4 (univariate regression), and figure J.5 (multiclass and multivariate). The regret figures report the mean rather than the median because the regret distribution is zero-inflated—a good metric picks the best model in many repeats, collapsing the median to zero.

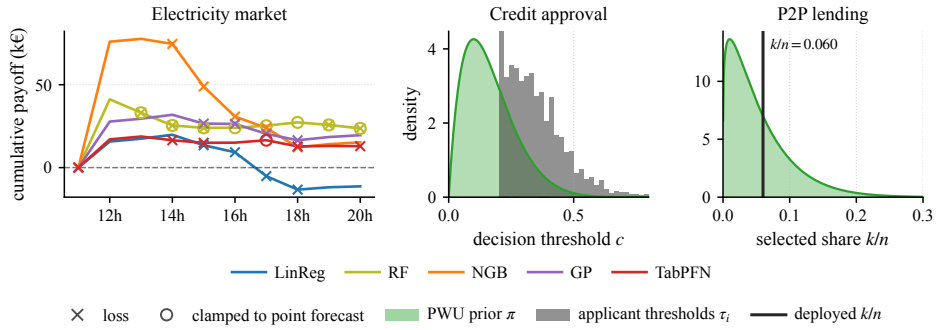


Figure J.9: Illustrations of our three case studies: cumulative electricity market trading profit on 08/08/2025 for 5 of our 10 models, showing when which models led to a conservative point-forecast bid (clamped) and which trades led to losses *(left)*, empirical distribution of the Bayes-optimal decision thresholds for the credit customers *(center)*, and the share of accepted loans with a fixed budget B *(right)*; both compared to our PWU prior choice. We can see that the credit thresholds differ significantly from our chosen prior, whereas our k/n-prior captures the real-world setup well; our PWU metrics achieve strong alignment in both cases (see figure J.10).

J.2 Applied case studies

Electricity market bidding We consider a wind farm operator who wants to sell electricity in the *day-ahead market*. In the day-ahead market, bids must be placed one day in advance, at noon, for each hour h of the next day. Hence, the wind farm operator needs a *forecasting model* for the next-day electricity generation E_h to decide on the bidding amounts $y_h, h = 0, \ldots, 23$. If a forecast is imprecise and the operator bids too much or too little electricity, they risk a penalty for the resulting shortfall or surplus, as, in both cases, the electricity grid's supply-demand balance needs to be restored. The operator therefore caps the expected imbalance risk, which ties how far the bid may depart from the point forecast to how uncertain that forecast is. In Bruninx et al. (2025), the optimization problem for the bidding amount y_h^* for hour h is defined as follows:

$$\begin{aligned} \max_{y_h} \quad & \mathbb{E}\left[R_h^{\mathrm{DA}} y_h + R_h^{\mathrm{B}}(E_h - y_h)\right] \\ \text{s.t.} \quad & \mathbb{E}\left[(E_h - y_h)^2\right] \leq \alpha, \\ & 0 \leq y_h \leq \beta, \end{aligned}$$

where β is the wind farm's capacity, E_h is the realized electricity generation, R_h^{DA} is the day-ahead market price (the unit payoff for the bid quantity y_h), and R_h^{B} is the balancing market price at hour h (the unit cost or reward for the deviation $E_h - y_h$). α is a pre-specified imbalance budget termed *risk certificate*, which we fix at $\alpha = 5000\,\mathrm{MWh}^2$. Bruninx et al. (2025) derive the optimal bidding amount y_h^* for hour h analytically. It is given as

$$y_h^* = \begin{cases} \min\{\mathbb{E}[E_h] + \Delta_h; \beta\} & \text{if } \mathbb{E}\left[R_h^{\mathrm{DA}}\right] > \mathbb{E}\left[R_h^{\mathrm{B}}\right], \\ \max\{\mathbb{E}[E_h] - \Delta_h; 0\} & \text{else}, \end{cases}$$

where $\Delta_h = \sqrt{\alpha - \mathrm{Var}[E_h]}$. A larger reported variance thus shrinks the admissible deviation Δ_h, pulling the bid towards the point forecast. Once $\mathrm{Var}[E_h] \geq \alpha$, the risk-minimizing bid is the point forecast itself, i.e. $\Delta_h = 0$. The per-hour payoff is then,

$$\left(R_h^{\mathrm{DA}}\, y_h^* + R_h^{\mathrm{B}}(E_h - y_h^*)\right),$$

where, following Bruninx et al. (2025), the operator is not a price taker: its open position $E_h - y_h^*$ shifts the system imbalance, and R_h^{B} is the marginal price of the balancing energy needed to restore the balance at that shifted volume.

As in Bruninx et al. (2025), we use Belgian balancing price data from Elia (2025) and day-ahead price and wind power data (*Belwind Phase 1* wind farm) from ENTSO-E (2025). The data from Elia

(2025) is licensed under CC BY 4.0. The data from ENTSO-E (2025) has no explicit license, but our use complies with their terms and conditions. We train on data from February to July 2025 to predict wind power generation, using the same suite of regression models as in our benchmark experiments. We then employ the trained models to estimate $\mathbb{E}[E_h]$ and $\mathrm{Var}[E_h]$ for each hour h in August 2025, based on which we compute y_h^* and the payoff. To quantify the variability of the metric–utility alignment scores, we bootstrap daily blocks (24 hours) of the 31 days of August, using $B = 100$ resamples. For each bootstrap replicate, we compute the metrics and utilities on the resampled set of days; the *bidding utility* is the sum of payoffs over all hours in the replicate,

$$\text{Bid-Util} = \sum_h \left(R_h^{\text{DA}}\, y_h^* + R_h^{\text{B}}(E_h - y_h^*) \right).$$

We visualize the case study on the left of figure J.9, and show the results in table 1 and on the left of figures J.10 and 4, and see that our PWU metrics exhibit the strongest utility-alignment, with PWU-SP also attaining the lowest realized regret.

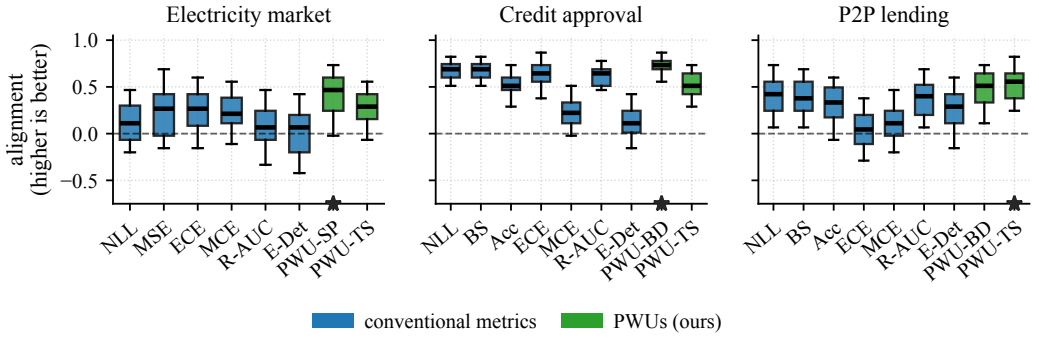


Figure J.10: Metric–utility alignment (Kendall's τ) in our real-world case studies. Each box's black line shows the median Kendall's τ value, the box spans the 25$^{\text{th}}$-75$^{\text{th}}$ percentile, and the whiskers the 5$^{\text{th}}$-95$^{\text{th}}$ percentile over 100 repeats. We mark the metric with the highest median with a star. Our PWU metrics exhibit strong alignment, even though our prior is misspecified for the credit case study, and the electricity market case study is not anticipated by any of our PWU metrics.

Credit approval In our second case study, we consider a dataset of credit requests. The decision-maker (the financial institution) needs to decide for each customer whether to grant or decline a requested credit. The modeling task is probabilistic binary classification, predicting the probability of default ($y = 1$). We follow the setup from Bahnsen et al. (2014), where the cost of a false negative (credit granted that defaults) is 75% of the credit amount, and the cost of a false positive (credit declined that would have been paid back) is the lost profit:

$$C_{\text{FN},i} = Cl_i \cdot 0.75, \quad C_{\text{FP},i} = r_i + C_{\text{FP}}^a,$$

where Cl_i is the customer's credit line, r_i is the loss in profit by rejecting a good customer, and $-C_{\text{FP}}^a$ is the gain through giving out the loan to an alternative customer. Cl_i and r_i are not readily available from the data, but Bahnsen et al. (2014) provide a way to calculate these quantities in appendix A of their paper. The resulting dataset-level utility is

$$-\sum_{i=1}^{n} \left[y_i C_{\text{FN},i} \mathbf{1}_{f_i \le \tau_i} + (1 - y_i) C_{\text{FP},i} \mathbf{1}_{f_i > \tau_i} \right],$$

where we choose the Bayes-optimal $\tau_i = \frac{C_{\text{FP},i}}{C_{\text{FP},i} + C_{\text{FN},i}}$ for each customer separately. We show a histogram of the τ_i in the center of figure J.9. The problem is similar to a binary decision, but the false negative and false positive costs are customer-dependent and highly variable. We use the Kaggle credit dataset (Cukierski, 2011), one of the two datasets of Bahnsen et al. (2014). It has no associated license, but our use respects the corresponding competition rules. To make the task nontrivial for our models, we further restrict to the marginal applicants, those with at least two recorded late payments

and revolving utilization above 50%. Without this restriction a single model is best in all 100 repeats and five metrics identify it every time, so realized decision-regret degenerates to a tie at zero and cannot separate the metrics.

We show the results in the center of figures J.10 and 4. The binary decision PWU has the strongest alignment with the credit approval utility of all nine metrics, and it also attains the lowest realized decision-regret, even though its prior is misspecified in light of the true τ-distribution.

Peer-to-peer lending In peer-to-peer (P2P) lending, the task is to assemble a loan portfolio from a large pool of borrowers, given a fixed budget B. We follow the setup from Byanjankar et al. (2021), who use the Lending Club dataset (George, 2020) that uses a CC0 1.0 Universal license. As in that setup, the lender has a fixed budget that we set to \$3000, and does not fund whole loans but buys fractions of them, subject to the Lending Club per-loan minimum of \$25, so the budget buys $B/25 = 120$ positions. Following Byanjankar et al. (2021), who evaluate on randomly drawn segments of 2,000 loans, we score each repeat–fold on an offering segment of 2,000 listings. The modeling task is probabilistic binary classification, predicting the probability of a good loan ($y = 1$). We greedily buy notes in the loans with the highest predicted probabilities until the budget B is exhausted. The utility is then computed as the return through the borrower on the amount lent for those loans that are paid back, minus the lent money for those loans that have not been paid back. To again make the task nontrivial for our models, we considered only loans with grade 6 or higher, corresponding to worse gradings. The task resembles top-k selection, but the payoff is customer-dependent and highly variable. Because every position has the same size, k/n is fixed at $120/2{,}000 = 6\%$ that we show on the right of figure J.9; so our PWU-TS prior choice is a good match. We show the results on the right of figures J.10 and 4. PWU-TS aligns best with the P2P lending utility and also attains the lowest realized regret.

Overall, our three case studies capture a scenario our PWUs reflect well (P2P lending), a scenario resembling a PWU decision family, but with misspecified prior (credit approval), and a scenario not anticipated at all (electricity market bidding). Our PWU metrics prove their suitability as UQ evaluation metric in all three scenarios.

J.3 Sensitivity analysis

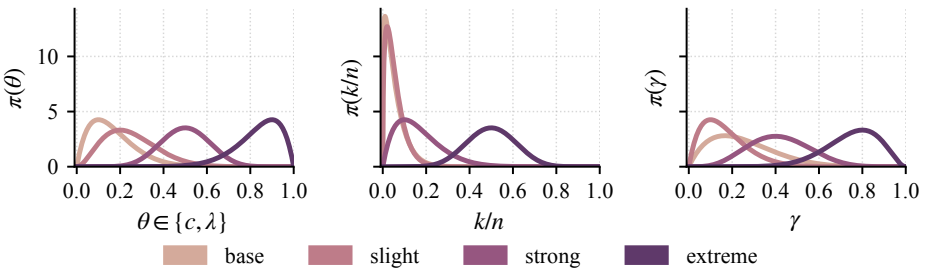


Figure J.11: Prior perturbations considered in our sensitivity analysis. The base priors anticipated by our PWU metrics are shown in the lightest color, while darker shades correspond to increasing levels of prior misspecification.

When constructing a PWU metric for a given decision family, one must commit to a specific prior. In this section, we analyze the sensitivity of PWU metrics to *prior misspecification*. We rerun a subset of our benchmark dataset experiments (binary classification, univariate regression, tabular data), but instead of measuring alignment to utilities U_θ originating from the same prior assumption π_θ, we measure the alignment to utilities stemming from perturbed priors $\tilde{\pi}_\theta$, so that our PWU metric priors are misspecified. We show the choice of our perturbed priors $\tilde{\pi}_\theta$ in figure J.11, alongside the *base prior* our PWU anticipates. We consider a *slight*, *strong*, and *extreme* perturbation. In figures J.12 and J.13, we show how the metric–utility alignments change across these perturbation levels in the binary classification and regression experiments respectively.

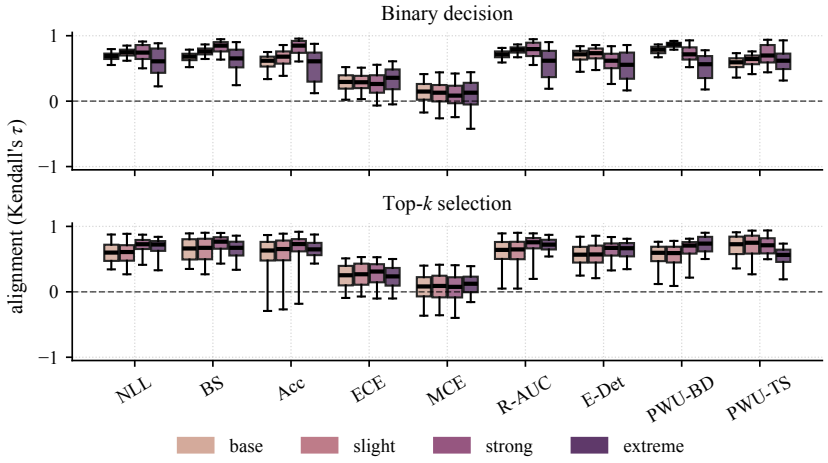


Figure J.12: Metric–utility alignment (Kendall's τ) in binary classification on tabular data under increasing levels of prior perturbation. Each box's black line shows the median Kendall's τ value, the box spans the 25$^{\text{th}}$-75$^{\text{th}}$ percentile, and the whiskers the 5$^{\text{th}}$-95$^{\text{th}}$ percentile over 100 repeats. The boxes corresponding to the base prior coincide with those shown in our benchmark experiments. Our PWU metrics exhibit strong robustness even under extreme prior misspecification.

Binary decision Our base prior for the binary decision problem has a mode at 10%, reflecting asymmetric costs in which false negatives are substantially more costly than false positives. The perturbed priors shift the mode to 20% (slight), 50% (strong), and 90% (extreme). The metric PWU-BD is the PWU metric that is decision-aligned with respect to the binary decision problem under the base prior. We can see that a slight perturbation hardly affects its utility-alignment, and the strong and extreme perturbation slightly worsens it. However, even after extreme perturbation, our PWU metric is still among the most aligned ones, demonstrating its robustness against prior-misspecification. Notably, even PWU-TS exhibits fair alignment across different priors. Another interesting observation is the development of Acc. Remember that Acc is decision-aligned with respect to the binary decision problem for $\pi_c = \delta_{0.5}$, so we would expect a strong alignment for the Beta prior with mode at 0.5—which is exactly what we can observe.

Top-k selection (classification) Our base prior for the top-k selection problem has a mode at 1%, reflecting the task of selecting a small number of items from a large pool. The perturbed priors shift the mode to 2% (slight), 10% (strong), and 50% (extreme). Slight and strong perturbation hardly affects the PWU-TS alignment. Extreme perturbations worsens its alignment, and conventional metrics take over. This makes sense since with increasing k/n, more datapoints of the datasets are considered in the top-k utility, and conventional metrics evaluate on the full dataset. However, even in this extreme regime, PWU-TS alignment is not far off the alignment of conventional metric, and remarkably, PWU-BD is among the most aligned ones.

Selective prediction Our base prior for selective prediction has a mode at 10%, while the perturbations consider the modes 20% (slight), 50% (strong), and 90% (extreme). PWU-SP is our PWU that is decision-aligned with respect to the selective prediction problem under the base prior. Slight and strong perturbation does not impair its performance. Even after extreme perturbation, our PWU metric is still among the most aligned ones, demonstrating its robustness against prior-misspecification. Another interesting observation is the development of the MSE. Remember that the MSE is decision-aligned for $\pi_\lambda = \delta_\infty$, so we would expect a strong alignment for the Beta prior with a high mode—which is exactly what we can observe.

Top-k selection (regression) Our base prior for the top-k selection problem has a mode at $k/n = 1\%$ and a moderate risk-aversion parameter γ. The perturbed priors consider $k/n = 2\%$ with low risk aversion (slight), $k/n = 10\%$ with strong risk aversion (strong), and $k/n = 50\%$ with extreme

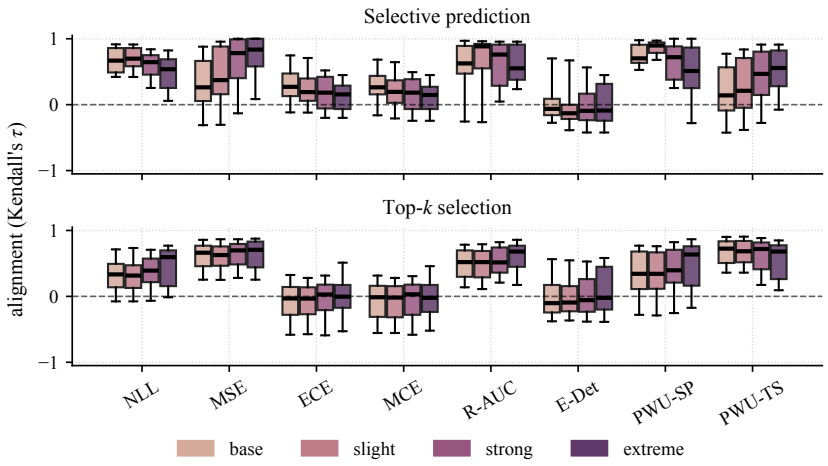


Figure J.13: Metric–utility alignment (Kendall's τ) in univariate regression on tabular data under increasing levels of prior perturbation. Each box's black line shows the median Kendall's τ value, the box spans the 25$^{\text{th}}$-75$^{\text{th}}$ percentile, and the whiskers the 5$^{\text{th}}$-95$^{\text{th}}$ percentile over 100 repeats. The boxes corresponding to the base prior coincide with those shown in our benchmark experiments. Our PWU metrics exhibit strong robustness even under extreme prior misspecification.

risk aversion (extreme). The metric PWU-TS is the PWU metric that is decision-aligned with respect to the top-k selection problem under the base prior. Slight and strong perturbations have essentially no effect on its utility alignment, while extreme perturbation leads to a moderate degradation in the lower percentiles. Nevertheless, even under extreme perturbation, PWU-TS remains among the most utility-aligned metrics, demonstrating robustness to prior misspecification.

Overall, these results show that PWU metrics are robust to prior misspecification, further underscoring their suitability in general-purpose UQ benchmarking.

Remark J.1 (Utility misspecification robustness). The sensitivity analysis above studies the robustness of a PWU metric when the prior π is misspecified *within* its target utility family. A complementary form of misspecification, however, occurs when the entire utility family is misspecified. Our experiments already provide empirical evidence on this scenario, since we evaluate every applicable PWU against every utility family considered. Inspecting figures 3, J.1, and J.2 we observe that most PWUs remain among the most decision-aligned metrics even when applied across utility families. A plausible explanation is that all of our PWUs are proper scoring rules and thus a fair heuristic in any application considered. Combined with the prior-level sensitivity analysis, this suggests that PWU metrics degrade gracefully under both forms of misspecification, further supporting their use as default metrics in general-purpose UQ benchmarking.

J.4 Compute infrastructure and PWU tractability

We ran all experiments on an internal high-performance compute cluster. Model training in the tabular benchmarks and case studies used GPU nodes, with each job allocated one NVIDIA A100 (40 GB MIG slice), 16 CPU cores, and 60 GB of memory. Metric evaluation—including all PWUs and the prior-sensitivity analyses, which only re-evaluate stored predictions under perturbed priors—ran on CPU nodes and takes less than one minute per repeat–fold combination (at most 30 s, in the image and text regression suites; a few seconds for the classification benchmarks and case studies).

Each PWU is a one- or two-dimensional integral of the utility over the prior, which we estimate by Monte Carlo with $S = 10^4$ prior draws. Some PWU metrics like the PWU-BD admit a closed form, but we used a unified MC scheme for all PWUs. On a single CPU core at $n = 2{,}000$ (the test size of our image and text repeats), one PWU evaluation takes between 0.3 ms and 0.9 s. The two-parameter regression PWU-TS is the most expensive at 0.9 s, since its ordering depends on the drawn shortfall penalty and must therefore be recomputed per draw; PWU-BD (0.07 s) and the

regression PWU-SP (0.12 s) fall in between. Even the slowest is orders of magnitude slower than closed-form metrics such as the NLL (< 1 ms), but negligible, since test-set evaluation only needs to be run once. To quantify the Monte Carlo error, we exploit that both classification PWUs admit closed forms, though of different type. For the PWU-BD, the integrand is piecewise linear in the cost c, so the integral reduces to truncated moments of the Beta prior. For the PWU-TS, the utility is a step function of the selection fraction, so the integral collapses to the finite sum $\sum_{k=1}^{n-1} -U_k(\boldsymbol{f}, \boldsymbol{y})\, \pi_k(k)$ over the discrete prior derived in appendix I.4. Across all 10 models of an image-classification fold ($n \in \{499, 2{,}000\}$), the $S = 10^4$ estimates deviate from the exact values by at most 1.1×10^{-3} in relative terms, the seed-to-seed standard deviation amounts to about 1% of the spread of PWU values across models, and the induced model ranking is identical across 50 independent Monte Carlo seeds.

K Omitted metric definitions

In this section, we define all metrics used in this paper and prove for each metric whether it is coordinate-independent or not. We start by defining each metric and giving a short intuition of why they are coordinate-(in)dependent. At the bottom of this section, we state a proposition, where we formally prove these intuitions. Throughout, ties are broken in favour of the lower instance index.

Negative log-likelihood (NLL) For binary classification, the pointwise NLL is defined as

$$l_{\text{NLL}}(f, y) = -y \log(f) - (1 - y) \log(1 - f),$$

where, for the NLL only, the prediction space is taken as $\mathbb{P} = (0, 1)$ (and, in the multiclass case of appendix F.2, as the predictions with $f^y > 0$), so that the NLL is real-valued as required by definition 3.1, and its integral representation in proposition 3.6 is finite. For regression ($f = (\mu, \sigma^2)$), we use the Gaussian NLL,

$$l_{\text{NLL}}(f, y) = \tfrac{1}{2} \log(2\pi\sigma^2) + \tfrac{(y-\mu)^2}{2\sigma^2}.$$

The dataset-level loss is defined as the average,

$$\text{NLL}(\boldsymbol{f}, \boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} l_{\text{NLL}}(f_i, y_i).$$

Thus, by definition, the **NLL is pointwise-separable, hence also coordinate-independent**.

Brier score (BS) The pointwise BS is defined as

$$l_{\text{BS}}(f, y) = (f - y)^2,$$

and the dataset-level loss is defined as the average,

$$\text{BS}(\boldsymbol{f}, \boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} l_{\text{BS}}(f_i, y_i).$$

Thus, by definition, the **BS is pointwise-separable, hence also coordinate-independent**.

Accuracy (Acc) Accuracy (Acc) is conventionally defined as the ratio of correctly classified instances when using a generic decision threshold of 0.5. To conform with our notion that smaller metrics correspond to better performance, we consider the negative conventional accuracy as our Acc. With this, it is the sample average of the pointwise decision 0-1 loss, the indicator function of a false classification of any kind, minus 1, i.e.,

$$l_{\text{0-1}}(f, y) = y \mathbf{1}_{f \leq 0.5} + (1 - y) \mathbf{1}_{f > 0.5},$$

$$\text{Acc}(\boldsymbol{p}, \boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} l_{\text{0-1}}(p_i, y_i) - 1.$$

Thus, by definition, **Acc is pointwise-separable, hence also coordinate-independent**.

Mean squared error (MSE) The pointwise mean squared error (MSE) is the squared-error loss, defined as

$$l_{\text{MSE}}(f, y) := (\mu - y)^2,$$

where $f = (\mu, \sigma^2)$, and the dataset MSE is

$$\text{MSE}(\boldsymbol{f}, \boldsymbol{y}) := \frac{1}{n} \sum_{i=1}^{n} l_{\text{MSE}}(f_i, y_i).$$

Thus, by definition, the **MSE is pointwise-separable, hence also coordinate-independent**. Note that the MSE only evaluates the predictive mean, not the full Gaussian predictive distribution.

Expected calibration error (ECE) A widely used empirical calibration metric in binary classification is the ECE, which measures the discrepancy between empirical frequencies and stated probabilities (aka *calibration*) within bins of the prediction space. Formally, fix a binning scheme $\mathcal{B} = \{B_1, \ldots, B_m\}$ that partitions $[0, 1]$. The ECE is defined as

$$\text{ECE}(\boldsymbol{f}, \boldsymbol{y}) := \sum_{j=1}^{m} \frac{n_j}{n} \big|\text{freq}(B_j) - \text{conf}(B_j)\big|,$$

where n_j is the number of indices i s.t. $f_i \in B_j$, $\text{freq}(B_j) = \frac{1}{n_j} \sum_{i:\, f_i \in B_j} y_i$, and $\text{conf}(B_j) = \frac{1}{n_j} \sum_{i:\, f_i \in B_j} f_i$. We use 10 equally spaced bins. The **ECE is coordinate-dependent**, since it couples coordinates through bin membership and bin statistics.

To use the ECE in regression, we follow Cui et al. (2020), who assess calibration of regression models through nominal coverage levels $\alpha_1, \ldots, \alpha_m \in (0, 1)$ and the corresponding Gaussian prediction intervals. We use $\alpha \in \{0.5, 0.8, 0.9, 0.95\}$. For Gaussian reports, the α_j-level interval is $I_{\alpha_j}(\mu_i, \sigma_i^2) := [\mu_i - z_{(1+\alpha_j)/2}\sigma_i,\ \mu_i + z_{(1+\alpha_j)/2}\sigma_i]$, and we let

$$\hat{\alpha}_j((\boldsymbol{\mu}, \boldsymbol{\sigma}^2), \boldsymbol{y}) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{y_i \in I_{\alpha_j}(\mu_i, \sigma_i^2)\}}$$

denote its empirical *coverage*, so that $\hat{\alpha}_j = \alpha_j$ for a perfectly calibrated report. Then, the empirical regression ECE is defined as

$$\text{ECE}((\boldsymbol{\mu}, \boldsymbol{\sigma}^2), \boldsymbol{y}) := \frac{1}{m} \sum_{j=1}^{m} \big|\hat{\alpha}_j((\boldsymbol{\mu}, \boldsymbol{\sigma}^2), \boldsymbol{y}) - \alpha_j\big|.$$

As in the binary classification case, the regression ECE is coordinate-dependent.

Maximum calibration error (MCE) The MCE measures the worst-case deviation between empirical frequencies and stated probabilities across bins of the prediction space. Using the same notation as for ECE, the MCE is defined as

$$\text{MCE}(\boldsymbol{f}, \boldsymbol{y}) := \max_{j \in \{1, \ldots, m\}} \big|\text{freq}(B_j) - \text{conf}(B_j)\big|.$$

Like the ECE, the **MCE is coordinate-dependent**. We adapt the MCE to regression in the same way as the ECE.

Area under the retention curve (R-AUC) The R-AUC (Malinin et al., 2021) measures the usefulness of a model's uncertainty estimates for selecting correct predictions. For varying certainty thresholds τ, the *retention curve* tracks the error rate of a model when only predicting on the subset with certainty at least τ. When uncertainty is informative, incorrect predictions tend to be rejected first, resulting in a steep decline of the retention curve. By integrating this curve, R-AUC yields a single scalar that reflects both overall predictive accuracy and the ability of uncertainty estimates to rank errors. Formally, let $s_i = s(f_i)$ denote a scalar uncertainty score for instance i, with larger values indicating higher uncertainty. We use the predictive entropy,

$$s_i = -f_i \log f_i - (1 - f_i) \log(1 - f_i).$$

Let π be a permutation of $\{1, \dots, n\}$ s.t. $s_{\pi(1)} \geq \cdots \geq s_{\pi(n)}$. For $r \in \{0, 1, \dots, n\}$, define the retained index set

$$R_r := \{\pi(r+1), \dots, \pi(n)\},$$

corresponding to retaining a fraction $\tau = |R_r|/n$ of predictions and deferring the remaining r most uncertain instances to an oracle. The area under the retention curve (R-AUC) is then defined as

$$\text{R-AUC}(\boldsymbol{f}, \boldsymbol{y}) := \frac{1}{n} \sum_{r=0}^{n-1} \frac{|R_r|}{n} \big(1 + \text{Acc}(\boldsymbol{f}_{i:i\in R_r}, \boldsymbol{y}_{i:i\in R_r})\big) = \frac{1}{n} \sum_{r=0}^{n-1} \frac{1}{n} \sum_{i \in R_r} \mathbf{1}_{f_i \neq y_i},$$

where $1 + \text{Acc}$ is the error rate—remember that our Acc is the negative conventional accuracy. The **R-AUC is coordinate-dependent**: changing a single prediction f_i affects not only the error term $\mathbf{1}_{f_i \neq y_i}$ associated with instance i, but also the ordering induced by the uncertainty scores $\{s_k\}_{k=1}^n$.

For regression models, we use the Gaussian predictive entropy as uncertainty score and the MSE as error rate. Like in binary classification, the regression R-AUC is coordinate-dependent.

Error detection (E-Det) E-Det (Bouvier et al., 2022) evaluates how well uncertainty can be used as a score to discriminate between correct and incorrect predictions. We define the hard predictions as $\hat{y}_i = \mathbf{1}_{f_i > 0.5}$ and the corresponding error indicators as $e_i = \mathbf{1}_{\hat{y}_i \neq y_i}$. Again writing s_i for the uncertainty score (where we use predictive entropy, as above), E-Det is defined as the classical area under the receiver operating characteristic curve (AUROC), using s_i as predictions for the labels e_i. To conform with our notion that smaller metrics correspond to better performance, we consider the negative conventional error detection as our E-Det. The **E-Det is coordinate-dependent** because of its ranking characteristic, similar to the R-AUC.

In regression, we again use the Gaussian predictive entropy as uncertainty score and define errors e_i as predictions whose relative errors $|\frac{y-\mu}{|y|+\varepsilon}|$ exceed 10% (using a small $\varepsilon > 0$ to ensure the relative error is well-defined).

Proposition K.1 (Coordinate-(in)dependence of common UQ evaluation metrics)**.** *The NLL, BS, Acc, and MSE are **coordinate-independent**, and the ECE, MCE, R-AUC, and E-Det are **coordinate-dependent**.*

Proof. The coordinate-independence of the NLL, BS, Acc, and MSE follows directly from their pointwise-separability by definition. We now prove coordinate-dependence of the remaining metrics.

ECE & MCE: We prove the claim for ECE; the claim for MCE then follows by the same construction (indeed, in the construction of steps 1–3, only one bin is populated, so ECE = MCE).

Step 1: choose a rational frequency inside a non-degenerate bin. Let $B = (\ell, u]$ be a bin of nonzero width, i.e., $u - \ell > 0$. Pick any rational number $q \in (\ell, u)$; write it as $q = k/t$ with integers $t \geq 2$ and $k \in \{0, 1, \dots, t\}$. (Existence is guaranteed since every non-degenerate interval contains rationals.) Set $n := t$ and fix labels $\boldsymbol{y}$ with exactly k ones and $t - k$ zeros; step 4 treats the dataset sizes for which no admissible q exists. Hence, for any prediction vector $\boldsymbol{f}$ for which *all* coordinates lie in B, we have

$$\text{freq}(B) = \frac{1}{t} \sum_{i=1}^{t} y_i = \frac{k}{t} = q.$$

Step 2: fix two values $a < b$ in the same bin. Choose any $a, b \in B$ with $a < b$ and with a and b sufficiently close to q so that the values defined below remain inside B. (Concretely: since B is an open neighborhood of q relative to $(\ell, u]$, there exists $\delta > 0$ s.t. $(q - \delta, q + \delta) \subseteq B$; choose $a, b \in (q - \delta, q + \delta)$ with $a < b$.) Fix an index $i \in \{1, \dots, t\}$ (any choice works).

Step 3: build two backgrounds that make the direction of change flip. Define $c := \frac{tq - a}{t-1}$ and $c' := \frac{tq - b}{t-1}$. By construction, if we set all coordinates of $\boldsymbol{f}$ except i equal to c (respectively c'), then the bin confidence equals q when $f_i = a$ (respectively $f_i = b$):

$$\frac{1}{t}\Big(a + \sum_{r \neq i} c\Big) = \frac{a + (t-1)c}{t} = \frac{a + tq - a}{t} = q, \quad \frac{1}{t}\Big(b + \sum_{r \neq i} c'\Big) = \frac{b + (t-1)c'}{t} = q.$$

Since a and b were chosen sufficiently close to q, the values c and c' are also close to q and thus lie in B as well (note that we can rewrite c, c' as perturbations of q: $c = q + \frac{q-a}{t-1}$, $c' = q + \frac{q-b}{t-1}$). Therefore, in both constructions, all predictions fall into the *same* bin B. Now define two prediction vectors:

$$\boldsymbol{f}_1 := (a, \underbrace{c, \ldots, c}_{t-1 \text{ times}}), \qquad \tilde{\boldsymbol{f}}_1 := (b, \underbrace{c, \ldots, c}_{t-1 \text{ times}}),$$

and

$$\boldsymbol{f}_2 := (a, \underbrace{c', \ldots, c'}_{t-1 \text{ times}}), \qquad \tilde{\boldsymbol{f}}_2 := (b, \underbrace{c', \ldots, c'}_{t-1 \text{ times}}),$$

where the distinguished coordinate is i (we suppress the explicit placement notation for readability). Since all coordinates lie in B, the ECE reduces to the single-bin absolute deviation:

$$\mathrm{ECE}_{\boldsymbol{y}}(\boldsymbol{f}) = \big|\mathrm{freq}(B) - \mathrm{conf}(B)\big|.$$

For $\boldsymbol{f}_1$ we have $\mathrm{conf}(B) = q = \mathrm{freq}(B)$, hence $\mathrm{ECE}_{\boldsymbol{y}}(\boldsymbol{f}_1) = 0$. But for $\tilde{\boldsymbol{f}}_1$ we have

$$\mathrm{conf}(B) = \frac{b + (t-1)c}{t} = \frac{b + (tq - a)}{t} = q + \frac{b-a}{t},$$

so

$$\mathrm{ECE}_{\boldsymbol{y}}(\tilde{\boldsymbol{f}}_1) = \left|q - \left(q + \frac{b-a}{t}\right)\right| = \frac{b-a}{t} > 0.$$

Thus, under the background $\boldsymbol{r} = (c, \ldots, c)$, changing f_i from a to b *increases* the ECE: $\mathrm{ECE}_{\boldsymbol{y}}\big((a, \boldsymbol{r})\big) < \mathrm{ECE}_{\boldsymbol{y}}\big((b, \boldsymbol{r})\big)$. Conversely, for $\tilde{\boldsymbol{f}}_2$ we again have perfect calibration in B by construction, hence $\mathrm{ECE}_{\boldsymbol{y}}(\tilde{\boldsymbol{f}}_2) = 0$, whereas for $\boldsymbol{f}_2$ we compute

$$\mathrm{conf}(B) = \frac{a + (t-1)c'}{t} = \frac{a + (tq - b)}{t} = q - \frac{b-a}{t},$$

so

$$\mathrm{ECE}_{\boldsymbol{y}}(\boldsymbol{f}_2) = \left|q - \left(q - \frac{b-a}{t}\right)\right| = \frac{b-a}{t} > 0.$$

Hence, under the background $\boldsymbol{r}' = (c', \ldots, c')$, changing f_i from a to b *decreases* the ECE: $\mathrm{ECE}_{\boldsymbol{y}}\big((a, \boldsymbol{r}')\big) > \mathrm{ECE}_{\boldsymbol{y}}\big((b, \boldsymbol{r}')\big)$.

Step 4: dataset sizes for which no admissible q exists. Steps 1–3 set $n := t$, the denominator of the chosen frequency q. For a *given* dataset size n, an admissible q fails to exist exactly when n divides the number of bins: every k/n is then an integer multiple of the bin width, hence a bin boundary rather than an interior point. With the 10 equally spaced bins of appendix K, this occurs only for $n \in \{2, 5, 10\}$. For such n, coordinate-dependence follows from a boundary-crossing construction instead, in which changing the distinguished coordinate moves it into or out of the bin occupied by the background. Let $y_1 = 1$ and $y_2 = 0$, and let $a := 0.45$ and $b := 0.55$, which lie in adjacent bins. For $n = 2$ and the background $r := 0.44$, the two coordinates share a bin under a but not under b, and

$$\mathrm{ECE}_{\boldsymbol{y}}\big((a, r)\big) = 0.055 < 0.445 = \mathrm{ECE}_{\boldsymbol{y}}\big((b, r)\big),$$

whereas under $r' := 0.56$ the roles are reversed and

$$\mathrm{ECE}_{\boldsymbol{y}}\big((a, r')\big) = 0.555 > 0.055 = \mathrm{ECE}_{\boldsymbol{y}}\big((b, r')\big).$$

The same four configurations give $0.055 < 0.45$ and $0.56 > 0.055$ for the MCE, so the direction of the effect of changing the first coordinate depends on the remaining one for both metrics. For $n \in \{5, 10\}$, set the remaining $n - 2$ coordinates to 0.01 with label 0: they occupy a bin disjoint from those of a, b, r, r', contribute the same amount to the ECE in all four configurations, and contribute a deviation of 0.01 to the MCE that is dominated by the two active bins. Both comparisons therefore persist, with the contributions of the active bins rescaled by $2/n$.

Transfer to regression. We give an explicit counterexample at the four levels $\alpha \in \{0.5, 0.8, 0.9, 0.95\}$ of appendix K, which settles the ECE and the MCE simultaneously. Let $n = 20$, $\boldsymbol{y} = \boldsymbol{0}$, and let every predictive variance equal 1. By appendix K, $y_i = 0$ is covered at level α exactly when $|\mu_i| \le z_{(1+\alpha)/2}$, where

$$z_{0.75} \approx 0.6745, \qquad z_{0.90} \approx 1.2816, \qquad z_{0.95} \approx 1.6449, \qquad z_{0.975} \approx 1.9600.$$

Fix the first coordinate and let a and b be the unit-variance reports with $\mu_1 = 0$ and $\mu_1 = 3$. Let the two backgrounds $\boldsymbol{r}, \boldsymbol{r}' \in \mathbb{P}^{19}$ also have unit variances, with predictive means

$$\boldsymbol{r}: \text{ nine } 0\text{s}, \text{ six } 1\text{s}, \text{ two } 1.5\text{s}, \text{ one } 1.8, \text{ one } 3, \qquad \boldsymbol{r}': \text{ ten } 0\text{s}, \text{ six } 1\text{s}, \text{ two } 1.5\text{s}, \text{ one } 1.8.$$

Counting the means that fall within each half-width gives the empirical coverage vectors $(\hat{\alpha}_j)_{j=1}^4$

	under a	under b
$\boldsymbol{r}$	$(0.50, 0.80, 0.90, 0.95)$	$(0.45, 0.75, 0.85, 0.90)$
$\boldsymbol{r}'$	$(0.55, 0.85, 0.95, 1.00)$	$(0.50, 0.80, 0.90, 0.95)$

Indeed, $|3|$ exceeds every half-width, so replacing a by b turns the first coordinate from covered into uncovered and lowers every coverage by exactly $1/20$; and $\boldsymbol{r}'$ differs from $\boldsymbol{r}$ only by replacing the mean 3 (uncovered at every level) with a mean 0 (covered at every level), which raises every coverage by $1/20$. Since $\text{ECE} = \frac{1}{4}\sum_j |\hat{\alpha}_j - \alpha_j|$ and $\text{MCE} = \max_j |\hat{\alpha}_j - \alpha_j|$, the table yields

$$\text{ECE}_{\boldsymbol{y}}\big((a, \boldsymbol{r})\big) = \text{MCE}_{\boldsymbol{y}}\big((a, \boldsymbol{r})\big) = 0, \qquad \text{ECE}_{\boldsymbol{y}}\big((b, \boldsymbol{r})\big) = \text{MCE}_{\boldsymbol{y}}\big((b, \boldsymbol{r})\big) = 0.05,$$

$$\text{ECE}_{\boldsymbol{y}}\big((a, \boldsymbol{r}')\big) = \text{MCE}_{\boldsymbol{y}}\big((a, \boldsymbol{r}')\big) = 0.05, \qquad \text{ECE}_{\boldsymbol{y}}\big((b, \boldsymbol{r}')\big) = \text{MCE}_{\boldsymbol{y}}\big((b, \boldsymbol{r}')\big) = 0.$$

Changing the first coordinate from a to b therefore *increases* both calibration errors under the background $\boldsymbol{r}$, but *decreases* both under $\boldsymbol{r}'$. Hence the regression ECE and MCE are coordinate-dependent.

R-AUC: We give explicit counterexamples that violate coordinate-independence. We start with the binary classification case. Let $n = 2$ and fix labels $\boldsymbol{y} = (1, 0)$. For a probability $f \in (0, 1)$, write $s(f) := -f \log f - (1 - f)\log(1 - f)$ for the predictive entropy, and let the hard prediction be $\hat{y}(p) := \mathbf{1}_{p>0.5}$. For $n = 2$ the R-AUC reduces to

$$\text{R-AUC}_{\boldsymbol{y}}(\boldsymbol{f}) = \frac{1}{2}\Big(\text{err}(\{1, 2\}) + \tfrac{1}{2}\,\text{err}(R_1)\Big),$$

where $\text{err}(\cdot)$ is the classification error rate, and R_1 is obtained by rejecting the single most uncertain instance. Fix the index $i = 1$ and choose $a := 0.9, b := 0.51$. Note that $\hat{y}(a) = \hat{y}(b) = 1$, hence the first instance is correct under both a and b, but $s(b) > s(a)$.

Background 1. Let the second prediction coordinate be $r := 0.8$. Then $\hat{y}(r) = 1 \neq y_2$, so the second instance is wrong, and moreover $s(b) > s(r) > s(a)$. Hence, with $\boldsymbol{f} = (a, r)$ the most uncertain instance is 2, so $R_1 = \{1\}$ and $\text{err}(R_1) = 0$; while with $\boldsymbol{f} = (b, r)$ the most uncertain instance is 1, so $R_1 = \{2\}$ and $\text{err}(R_1) = 1$. In both cases the full-set error rate is $\text{err}(\{1, 2\}) = 1/2$ (exactly one error). Therefore,

$$\text{R-AUC}_{\boldsymbol{y}}((a, r)) = \frac{1}{2}\Big(\frac{1}{2} + \frac{1}{2}\cdot 0\Big) = \frac{1}{4} < \frac{1}{2} = \frac{1}{2}\Big(\frac{1}{2} + \frac{1}{2}\cdot 1\Big) = \text{R-AUC}_{\boldsymbol{y}}((b, r)),$$

so changing f_1 from a to b *increases* the R-AUC.

Background 2. Let the second prediction coordinate be $r' := 0.01$. Then $\hat{y}(r') = 0 = y_2$, so the second instance is correct, and $s(r') < s(a) < s(b)$. Hence, regardless of whether $f_1 = a$ or $f_1 = b$, the most uncertain instance is 1 and thus $R_1 = \{2\}$, yielding $\text{err}(R_1) = 0$. Moreover, both instances are correct under (a, r') and (b, r'), so $\text{err}(\{1, 2\}) = 0$ in both cases. Thus,

$$\text{R-AUC}_{\boldsymbol{y}}((a, r')) = \text{R-AUC}_{\boldsymbol{y}}((b, r')) = 0.$$

Combining the two backgrounds, we have exhibited fixed labels $\boldsymbol{y}$, an index $i = 1$, and values a, b s.t.

$$\text{R-AUC}_{\boldsymbol{y}}((a, r)) < \text{R-AUC}_{\boldsymbol{y}}((b, r)), \;\; \text{while} \;\; \text{R-AUC}_{\boldsymbol{y}}((a, r')) = \text{R-AUC}_{\boldsymbol{y}}((b, r')),$$

violating coordinate-independence.

Now we move on to regression. Let $n = 2$ and fix labels $\boldsymbol{y} = (0, 0)$. Reports are $f_i = (\mu_i, \sigma_i^2)$, the uncertainty score is the Gaussian predictive entropy $s(\mu, \sigma^2) := \frac{1}{2}\log(2\pi e\,\sigma^2)$, which is strictly increasing in σ^2, and the error rate on a retained set is the MSE computed from the predictive means. Fix $i = 1$ and choose the two alternative reports

$$a := (\mu_1, \sigma_1^2) = (1, 0.1^2), \quad b := (\mu_1, \sigma_1^2) = (1, 10^2),$$

so that the mean-squared error of the first instance equals $(\mu_1 - y_1)^2 = 1$ under both a and b, but the uncertainty score satisfies $s(b) > s(a)$.

Background 1. Let the remaining coordinate be $r := (\mu_2, \sigma_2^2) = (0, 1^2)$, so the second instance has squared error $(\mu_2 - y_2)^2 = 0$ and uncertainty score $s(r)$ with $s(a) < s(r) < s(b)$. Hence, with (a, r) the most uncertain instance is 2 and $R_1 = \{1\}$, whereas with (b, r) the most uncertain instance is 1 and $R_1 = \{2\}$. The full-set MSE equals

$$\mathrm{MSE}(\{1, 2\}) = \frac{1 + 0}{2} = \frac{1}{2}$$

for both (a, r) and (b, r), while the retained-set MSE equals 1 under (a, r) and 0 under (b, r). Therefore,

$$\text{R-AUC}_{\boldsymbol{y}}((a, r)) = \frac{1}{2}\Big(\frac{1}{2} + \frac{1}{2} \cdot 1\Big) = \frac{1}{2} > \frac{1}{4} = \frac{1}{2}\Big(\frac{1}{2} + \frac{1}{2} \cdot 0\Big) = \text{R-AUC}_{\boldsymbol{y}}((b, r)),$$

so changing f_1 from a to b *decreases* the regression R-AUC.

Background 2. Let the remaining coordinate be $r' := (\mu_2, \sigma_2^2) = (0, 0.01^2)$, so $s(r') < s(a) < s(b)$. Then the most uncertain instance is 1 under both (a, r') and (b, r'), hence $R_1 = \{2\}$ in both cases. Again $\mathrm{MSE}(\{1, 2\}) = 1/2$, and the retained-set MSE is 0 in both cases, so

$$\text{R-AUC}_{\boldsymbol{y}}((a, r')) = \text{R-AUC}_{\boldsymbol{y}}((b, r')) = \frac{1}{4}.$$

Thus, the strict comparison between $\text{R-AUC}_{\boldsymbol{y}}((a, \cdot))$ and $\text{R-AUC}_{\boldsymbol{y}}((b, \cdot))$ depends on the remaining coordinate, proving that the regression R-AUC is also coordinate-dependent.

E-Det: Since E-Det is the negative conventional AUROC, it suffices to show that the conventional E-Det is coordinate-dependent; multiplying by -1 preserves coordinate-(in)dependence. We start with binary classification. Let $n = 3$, fix labels $\boldsymbol{y} = (1, 0, 1)$, and let $\hat{y}(f) := \mathbf{1}_{f > 0.5}$. Define the error indicators $e_i = \mathbf{1}_{\hat{y}(f_i) \neq y_i}$, and let $\mathrm{AUROC}(\boldsymbol{s}, \boldsymbol{e})$ be the standard AUROC treating larger s_i as stronger evidence for $e_i = 1$. Fix $i = 1$ and choose two values $a := 0.9, b := 0.49$. Then $\hat{y}(a) = 1$ (hence $e_1 = 0$ under a) while $\hat{y}(b) = 0$ (hence $e_1 = 1$ under b), and moreover $s(b) > s(a)$ since predictive entropy is maximized at 0.5. We construct two backgrounds $\boldsymbol{r}, \boldsymbol{r}' \in (0, 1)^2$ (for coordinates $2, 3$) that yield different directions of change.

Background 1. Set $(f_2, f_3) = (0.999, 0.6)$. Then $\hat{y}(f_2) = 1 \neq y_2 = 0$, so $e_2 = 1$, and $\hat{y}(f_3) = 1 = y_3$, so $e_3 = 0$. Moreover, $s(0.999) < s(0.9) < s(0.6) < s(0.49)$. Hence, under $(a, \boldsymbol{r})$ the positive set is $\{2\}$ and the negative set is $\{1, 3\}$, and the unique positive score $s_2 = s(0.999)$ is smaller than both negative scores, so $\text{AUROC}_{\boldsymbol{y}}((a, \boldsymbol{r})) = 0$. Under $(b, \boldsymbol{r})$ the positive set is $\{1, 2\}$ and the negative set is $\{3\}$, and we have $s(b) > s(0.6) > s(0.999)$, hence exactly one of the two positive-negative comparisons is correct, so $\text{AUROC}_{\boldsymbol{y}}((b, \boldsymbol{r})) = \frac{1}{2}$. Thus, changing f_1 from a to b *increases* AUROC under background $\boldsymbol{r}$.

Background 2. Set $(f_2, f_3) = (0.51, 0.99)$. Then $\hat{y}(f_2) = 1 \neq y_2 = 0$, so $e_2 = 1$, and $\hat{y}(f_3) = 1 = y_3$, so $e_3 = 0$. Moreover, $s(0.51)$ is close to maximal and therefore larger than both $s(0.9)$ and $s(0.99)$. Hence, under $(a, \boldsymbol{r}')$ the positive set is $\{2\}$ and negatives are $\{1, 3\}$, and s_2 exceeds both negative scores, so $\text{AUROC}_{\boldsymbol{y}}((a, \boldsymbol{r}')) = 1$. Under $(b, \boldsymbol{r}')$ the positives are $\{1, 2\}$ and the negative is $\{3\}$, and both positive scores $s(b)$ and $s(0.51)$ exceed the negative score $s(0.99)$, so again $\text{AUROC}_{\boldsymbol{y}}((b, \boldsymbol{r}')) = 1$. Thus, under background $\boldsymbol{r}'$, the same change from a to b leaves AUROC unchanged, violating coordinate-independence.

We now continue with the regression case. Let $n = 3$ and fix labels $\boldsymbol{y} = (1, 1, 1)$. Reports are $f_i = (\mu_i, \sigma_i^2)$. Fix $\tau > 0$ and $\varepsilon > 0$, and define error indicators $e_i := \mathbf{1}_{\left\{\left|\frac{y_i - \mu_i}{|y_i| + \varepsilon}\right| > \tau\right\}}$. Define the conventional AUROC using scores $s_i = s(f_i)$ to predict labels e_i. Fix $i = 1$ and choose two reports

$$a := (\mu_1, \sigma_1^2) = (1,\ 0.1^2), \quad b := (\mu_1, \sigma_1^2) = (1 + 2(|1| + \varepsilon)\tau,\ 10^2).$$

Then under a the relative error is 0, so $e_1 = 0$, while under b it is $2\tau > \tau$, so $e_1 = 1$, and $s(b) > s(a)$.

Background 1. Set

$$f_2 = (1 + 2(|1| + \varepsilon)\tau,\ 0.01^2), \quad f_3 = (1,\ 1^2).$$

Then $e_2 = 1$ and $e_3 = 0$. Moreover, by monotonicity in σ^2, $s(f_2) < s(a) < s(f_3) < s(b)$. Hence, under $(a, \boldsymbol{r})$ the positive set is $\{2\}$ and negatives are $\{1, 3\}$, and the unique positive score is smaller

than both negative scores, giving $\text{AUROC}_{\boldsymbol{y}}((a, \boldsymbol{r})) = 0$. Under $(b, \boldsymbol{r})$ the positives are $\{1, 2\}$ and the negative is $\{3\}$, and exactly one of the two positive-negative comparisons is correct (since $s(b) > s(f_3) > s(f_2)$), giving $\text{AUROC}_{\boldsymbol{y}}((b, \boldsymbol{r})) = \frac{1}{2}$. Thus, AUROC increases under background $\boldsymbol{r}$ when changing f_1 from a to b.

Background 2. Set

$$f_2 = (1 + 2(|1| + \varepsilon)\tau,\ 10^2), \quad f_3 = (1,\ 0.01^2).$$

Then $e_2 = 1$ and $e_3 = 0$, and $s(f_2)$ is large while $s(f_3)$ is small. Consequently, under $(a, \boldsymbol{r}')$ the positive set is $\{2\}$ and negatives are $\{1, 3\}$, and the positive score exceeds both negative scores, so $\text{AUROC}_{\boldsymbol{y}}((a, \boldsymbol{r}')) = 1$. Under $(b, \boldsymbol{r}')$ the positives are $\{1, 2\}$ and the negative is $\{3\}$, and both positive scores exceed the negative score, so again $\text{AUROC}_{\boldsymbol{y}}((b, \boldsymbol{r}')) = 1$. Thus, under background $\boldsymbol{r}'$, the same change from a to b leaves AUROC unchanged, violating coordinate-independence. □