What Does It Take to Build a Performant Selective Classifier?

Stephan Rabanser*
Princeton University
rabanser@princeton.edu

Nicolas Papernot
University of Toronto & Vector Institute
nicolas.papernot@utoronto.ca

Abstract

Selective classifiers improve model reliability by abstaining on inputs the model deems uncertain. However, few practical approaches achieve the gold-standard performance of a perfect-ordering oracle that accepts examples exactly in order of correctness. Our work formalizes this shortfall as the *selective-classification gap* and present the first finite-sample decomposition of this gap to five distinct sources of looseness: Bayes noise, approximation error, ranking error, statistical noise, and implementation- or shift-induced slack. Crucially, our analysis reveals that monotone post-hoc calibration—often believed to strengthen selective classifiers—has limited impact on closing this gap, since it rarely alters the model's underlying score ranking. Bridging the gap therefore requires scoring mechanisms that can effectively reorder predictions rather than merely rescale them. We validate our decomposition on synthetic two-moons data and on real-world vision and language benchmarks, isolating each error component through controlled experiments. Our results confirm that (i) Bayes noise and limited model capacity can account for substantial gaps, (ii) only richer, feature-aware calibrators meaningfully improve score ordering, and (iii) data shift introduces a separate slack that demands distributionally robust training. Together, our decomposition yields a quantitative error budget as well as actionable design guidelines that practitioners can use to build selective classifiers which approximate ideal oracle behavior more closely.

1 Introduction

In high-stakes applications like finance [Coenen et al., 2020], healthcare [Guan et al., 2020], and autonomous driving [Ghodsi et al., 2021], machine learning (ML) models are increasingly tasked with making decisions under uncertainty, where dependable predictions are critical. Selective classifiers [Chow, 1957, El-Yaniv et al., 2010] formalize the option to abstain on inputs deemed unreliable, reducing the risk of costly errors by refusing to predict when uncertain. Their effectiveness depends on identifying which predictions to trust and which to defer. A common evaluation metric is the *accuracy–coverage* tradeoff, which quantifies how performance degrades as the model accepts a broader set of inputs. The benchmark is a hypothetical oracle that ranks inputs by their true likelihood of correctness, yielding a *perfect-ordering upper bound* [Geifman et al., 2019, Rabanser et al., 2023]. While some selective predictors approach this bound, others fall short—revealing persistent gaps and raising open questions about what properties of the learning setup truly govern selective performance.

Classical theory explains selective classification in two idealized regimes. In the *realizable* setting [El-Yaniv et al., 2010], where the data is noiseless and the true predictor lies within the hypothesis class, the model can asymptotically achieve the ideal accuracy–coverage curve. In the more general *agnostic* setting [Wiener and El-Yaniv, 2011], the classifier competes with the best-in-class predictor, but this benchmark may itself fall well below the oracle bound—and the theory does not isolate

*Work done while at the University of Toronto and the Vector Institute.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

the source of the gap. Yet in practice, we never operate in such asymptotic or idealized conditions: models are often misspecified, the data used for training and evaluation are finite, and asymptotic guarantees offer little actionable insight. As a result, even the strongest formal guarantees provide limited guidance, which leaves practitioners with the following question:

For my finite model on finite data, what aspects of the learning setup will actually move my trade-off curve closer to the perfect-ordering upper bound?

To answer this question, we re-frame selective performance around the *selective classification gap* $\Delta(c)$: the mismatch between a model's accuracy–coverage curve and the oracle bound for all coverage levels c (see Figure 1). Our work shows that this gap admits a finite-sample decomposition:

$$\widehat{\Delta}(c) \;\le\; \underbrace{\varepsilon_{\text{Bayes}}(c)}_{\text{irreducible}} + \underbrace{\varepsilon_{\text{approx}}(c)}_{\text{capacity}} + \underbrace{\varepsilon_{\text{rank}}(c)}_{\text{ranking}} + \underbrace{\varepsilon_{\text{stat}}(c)}_{\text{data}} + \underbrace{\varepsilon_{\text{misc}}(c)}_{\text{optimization \& shift}}, \qquad \forall c \in (0,1]. \tag{1}$$

Each term corresponds to a distinct—and often *measurable*—source of looseness. The first term, $\varepsilon_{\text{Bayes}}(c)$, reflects irreducible uncertainty: if the true label is inherently unpredictable from the input (e.g., due to label noise), even a perfect classifier must abstain on some examples. Next, $\varepsilon_{\text{approx}}(c)$ captures limits of the model class: if the function class is too weak to approximate the Bayes-optimal decision rule, the gap widens. The third term, $\varepsilon_{\text{rank}}(c)$, quantifies the model's failure to correctly order inputs by their likelihood of correctness—typically due to poor confidence estimation or miscalibration. The statistical term $\varepsilon_{\text{stat}}(c)$ accounts for finite-sample fluctuations that affect both learning and evaluation. Finally, $\varepsilon_{\text{misc}}(c)$ aggregates practical imperfections, such as optimization error or test-time distribution shift. Equation (1) thus provides a coverage-uniform "error budget" that transforms the qualitative question posed earlier into a concrete quantitative diagnosis.

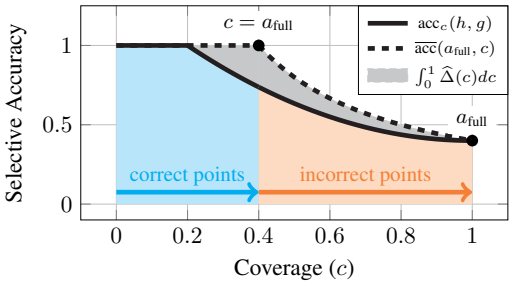


Figure 1: **Visualization of the selective classification gap $\Delta(c)$.** The dashed curve is the oracle frontier $\overline{\text{acc}}(a_{\text{full}}, c)$ under which coverage levels left of $c = a_{\text{full}}$ (blue) accept all correct predictions first, and rank incorrect predictions last (orange). This constitutes the ideal behavior of a selective predictor. On the other hand, the solid curve shows the realized selective accuracy $\text{acc}_c(h, g)$. The mismatch between $\overline{\text{acc}}(a_{\text{full}}, c)$ and $\text{acc}_c(h, g)$ at coverage c is the gap $\Delta(c)$; the gray shaded area visualizes this gap over the full coverage spectrum.

Two key insights, developed further in later sections, are worth previewing. First, we show that monotone post-hoc calibration—a technique that is often thought to improve selective prediction performance—only possess a limited ability of reducing the *ranking* term $\varepsilon_{\text{rank}}(c)$. In contrast, methods that directly yield improved uncertainty scores by leveraging richer feature representations or aggregating diverse model perspectives dominate post-hoc calibration methods. Second, Equation (1) serves as an *error budget* that identifies cost-effective levers leading to actionable recommendations for practitioners: (i) use additional or repeated labels and noise-robust losses to reduce $\varepsilon_{\text{Bayes}}$; (ii) increase capacity or distill from a more expressive teacher to shrink $\varepsilon_{\text{approx}}$; (iii) enlarge validation data to lower $\varepsilon_{\text{stat}}$; and (iv) apply domain adaptation or importance weighting to address $\varepsilon_{\text{misc}}$.

Contributions. We summarize our main contributions below:

- **Problem formulation.** We recast selective prediction in terms of a *coverage-uniform selective classification gap*—the key quantity to minimize to approach perfect selective prediction. This framing unifies prior work and highlights which failure modes dominate at each coverage level.
- **Theoretical analysis.** We present the first *finite-sample decomposition* of the selective classification gap (Equation (1)), dividing it into five terms: Bayes, approximation, ranking, statistical, and miscellaneous errors. Our analysis further shows that *monotone calibration is ineffective at reducing the gap*, motivating the use of methods that can change the ranking more flexibly.
- **Empirical validation.** Our synthetic and real-world experiments confirm the decomposition: Bayes noise and capacity limits drive large gaps; temperature scaling improves calibration but not ranking; and shift-aware methods remain essential under distribution shift. *These results clarify which factors matter most and how to target them effectively in practice.*

2 Background & Related Work on Selective Classification

Selective classification extends the standard supervised classification framework as follows:

Definition 1 (Selective Classifier [Chow, 1957, El-Yaniv et al., 2010])**.** A selective classifier is a pair (h, g), where $h : \mathcal{X} \to \mathcal{Y}$ is a classifier over covariates $\mathcal{X} = \mathbb{R}^D$ and labels $\mathcal{Y} = \{1, \ldots, K\}$, and $g : \mathcal{X} \times (\mathcal{X} \to \mathcal{Y}) \to \mathbb{R}$ is a selection function that assigns a confidence score. Given a threshold $\tau \in \mathbb{R}$, the model abstains when the score falls below the threshold:

$$(h, g)(x) = \begin{cases} h(x) & \text{if } g(x, h) \geq \tau \\ \perp & \text{otherwise} \end{cases}. \tag{2}$$

Intuitively, a selective classifier predicts only when confident. The selection score $g(x, h)$ determines whether to accept or abstain: if $g(x, h) \geq \tau$, the model outputs $h(x)$; otherwise, it returns $\perp$.

Many prior works have developed selective classification methods for training competitive pairs (h, g). A popular method is *Maximum Softmax Probability* (`MSP`) [Hendrycks and Gimpel, 2017, Geifman and El-Yaniv, 2017], which uses classifier confidence as the selection score. To improve calibration and reduce predictive variance, ensembling approaches have been explored: *Deep Ensembles* (`DE`) [Lakshminarayanan et al., 2017] train multiple models with different initializations, while *Selective Classification via Training Dynamics* (`SCTD`) [Rabanser et al., 2022] ensembles intermediate checkpoints. Other methods—such as *SelectiveNet* (`SN`)[Geifman and El-Yaniv, 2019], *Deep Gamblers* (`DG`)[Liu et al., 2019], and *Self-Adaptive Training* (`SAT`) [Huang et al., 2020]—alter the model architecture or loss function ensuring that prediction and rejection are learned jointly.

The efficacy of a selective classifier is evaluated using the empirical accuracy-coverage tradeoff.

Definition 2 (Empirical Accuracy–Coverage Tradeoff)**.** Let $D = \{(x_i, y_i)\}_{i=1}^N$ be a dataset. For a selective classifier (h, g) and threshold τ, define

$$\hat{\xi}_{h,g}(\tau) = \frac{1}{N} \left| \{ i : g(x_i, h) \geq \tau \} \right|, \tag{3}$$

$$\hat{\alpha}_{h,g}(\tau) = \begin{cases} \dfrac{\left| \{ i : h(x_i) = y_i \text{ and } g(x_i, h) \geq \tau \} \right|}{\left| \{ i : g(x_i, h) \geq \tau \} \right|}, & \text{if } \hat{\xi}_{h,g}(\tau) > 0, \\ 0, & \text{if } \hat{\xi}_{h,g}(\tau) = 0. \end{cases} \tag{4}$$

The pair $(\hat{\xi}, \hat{\alpha})$ as τ varies is the empirical accuracy–coverage curve.

The score $g(x, h)$ therefore induces a total order over D: x_1 is accepted before x_2 if $g(x_1, h) > g(x_2, h)$. This ordering governs which inputs are retained as coverage decreases. Effective strategies aim to maximize $\hat{\alpha}$ at each coverage level $\hat{\xi}$, often trading off accuracy and coverage.

Accuracy–coverage tradeoff evaluation. The accuracy–coverage tradeoff is often summarized by the area under the accuracy–coverage curve (`AUACC`), integrating selective accuracy over all coverage levels [Traub et al., 2024]. However, Geifman et al. [2019] show that `AUACC` favors models already accurate at full coverage. To address this issue, Geifman et al. [2019] and Rabanser et al. [2023] propose oracle-based bounds, which become loose at low utility [Galil et al., 2023]. To avoid accuracy bias, Galil et al. [2023] and Pugnana and Ruggieri [2023] recommend using the classifier's `AUROC` instead. But `AUROC` is not monotonic in `AUACC` [Cattelan and Silva, 2023, Ding et al., 2020], thus favoring methods tuned for `AUROC` over selective accuracy. Recently, Traub et al. [2024] introduced the Area Under the Goals-Reweighted Curve (`AUGRC`), which multiplies accuracy by coverage to mitigate bias toward low-coverage regions, while Mucsányi et al. [2024] provide a benchmark disentangling uncertainty sources for fairer comparison. These efforts refine *evaluation metrics*, whereas our work complements them by analyzing *what causes* selective performance gaps. Earlier work [El-Yaniv et al., 2010, Wiener and El-Yaniv, 2011] characterizes optimal selective classifiers in both realizable and agnostic regimes but focuses on existence rather than practical instantiation—unlike our finite-sample perspective.

3 Decomposing the Selective Classification Gap

We characterize the optimal performance achievable by a selective classifier given its full-coverage accuracy, establishing a reference against which all practical selective classifiers can be evaluated.

3.1 Oracle Bound and Selective Classification Gap

Definition 3 (Perfect Ordering Upper Bound [Geifman et al., 2019, Rabanser et al., 2023])**.** Fix a base classifier h whose *full-coverage* (standard) accuracy is $a_{\text{full}} := \Pr\big(h(X) = Y\big) \in [0, 1]$. For any desired coverage level $c \in (0, 1]$, the best selective accuracy—achieved by accepting the c-fraction of points with the *highest* posterior correctness $\Pr(h(X) = Y \mid X)$—is

$$\overline{\text{acc}}\big(a_{\text{full}}, c\big) = \begin{cases} 1, & 0 < c \leq a_{\text{full}}, \\ \dfrac{a_{\text{full}}}{c}, & a_{\text{full}} < c < 1. \end{cases} \tag{5}$$

Assuming no Bayes noise—that is, all errors are avoidable given perfect confidence—this piecewise curve (see Figure 1) traces the *oracle* accuracy–coverage frontier based on a perfect ranking of examples by correctness probability. Any real selective classifier falls below this bound—potentially far below, depending on its calibration, expressivity, and sensitivity to noise. To quantify how far a given classifier falls short of this ideal, we define the *selective classification gap*.

Definition 4 (Selective Classification Gap)**.** Let (h, g) be a selective classifier with full-coverage accuracy $a_{\text{full}} = \Pr(h(X) = Y)$. For a coverage level $c \in (0, 1]$, let τ_c be the threshold satisfying $\Pr\big(g(X, h) \geq \tau_c\big) = c$. The *selective accuracy* at coverage c is $\text{acc}_c(h, g) := \Pr\big(h(X) = Y \;\big|\; g(X, h) \geq \tau_c\big)$. The *selective classification gap* at coverage c is then defined as the deviation from the perfect-ordering upper bound:

$$\Delta(c) := \overline{\text{acc}}\big(a_{\text{full}}, c\big) \; - \; \text{acc}_c(h, g). \tag{6}$$

This gap $\Delta(c)$ can be interpreted as the *excess selective risk* at a given coverage c. We note that integrating $\Delta(c)$ over the entire coverage spectrum, $\int_0^1 \Delta(c)dc$, is equivalent to the definition of the Excess-AURC (E-AURC) metric proposed by Geifman et al. [2019].

The function $\Delta(c)$ offers a coverage-resolved diagnostic of selective performance. A small gap indicates near-oracle behavior—accepting only examples it can confidently and correctly classify—while a large gap suggests limitations in estimating correctness or ranking examples reliably. Understanding the magnitude and shape of this gap is key to analyzing and improving selective classifiers.

3.2 Why Is the Upper Bound Loose?

The oracle bound in Definition 3 relies on two idealized assumptions: perfect prediction on all inputs and perfect ranking by the true correctness posterior. In practice, selective classifiers deviate in four principal ways, each corresponding to a term in our later decomposition ($\varepsilon_{\text{Bayes}}, \varepsilon_{\text{approx}}, \varepsilon_{\text{rank}}, \varepsilon_{\text{stat}}$):

1. **Bayes noise ($\varepsilon_{\text{Bayes}}$).** Even a Bayes-optimal rule errs on intrinsically ambiguous points (where $\max_y \Pr(Y = y \mid X) < 1$), unavoidable in real data [Devroye et al., 2013]. As coverage increases, the oracle must accept some of these noisy inputs, lowering the achievable accuracy.
2. **Approximation limits ($\varepsilon_{\text{approx}}$).** A learned model h drawn from a restricted hypothesis class may misclassify inputs with high posterior confidence under the Bayes rule [Bishop, 2006]. This gap reduces full-coverage accuracy and limits selective performance.
3. **Ranking error ($\varepsilon_{\text{rank}}$).** Let $\eta_h(x) := \Pr\big(h(x) = Y \mid X = x\big)$ denote the true correctness posterior, i.e., the probability that the model's prediction is correct given the input. Ideally, the confidence score $g(X, h)$ should rank examples in decreasing order of $\eta_h(x)$—so that samples the model is likely to classify correctly (high $\eta_h(x)$, examples that are "easy") are accepted before those it is likely to misclassify (low $\eta_h(x)$, examples that are "hard"). When $g(X, h)$ fails to preserve this ordering, high-confidence errors and low-confidence corrects are interleaved, increasing the selective gap $\Delta(c)$.

4. **Statistical noise ($\varepsilon_{\text{stat}}$).** Estimating the threshold τ_c and selective accuracy from a finite validation set introduces randomness of order $\mathcal{O}(\sqrt{\log(1/\delta)/n})$. This follows from concentration bounds; see Shalev-Shwartz and Ben-David [2014] for standard applications in learning theory.

> **Takeaway.** The selective classification gap $\Delta(c)$ reflects a mix of irreducible noise, model misspecification, ranking errors, and sampling variability. Addressing each—via cleaner labels, stronger models, or improved ranking—can tighten selective prediction performance.

In the next subsection, we formalize this decomposition and provide a general bound on the total gap.

3.3 Formal Decomposition of the Gap

We now give a principled decomposition of the selective classification gap and provide a corresponding finite-sample upper bound. For clarity and notational simplicity, we treat the binary-label case $\mathcal{Y} = \{0, 1\}$; the multiclass extension follows by a standard one-vs-rest reduction.

Notation. Let $\eta(x) := \Pr\big(Y = 1 \mid X = x\big)$ be the Bayes posterior. For a fixed classifier $h : \mathcal{X} \to \mathcal{Y}$ define its (induced) correctness posterior

$$\eta_h(x) \;:=\; \Pr\big(h(x) = Y \mid X = x\big) \;=\; \eta(x)\,\mathbb{I}_{\{h(x)=1\}} + \big(1 - \eta(x)\big)\mathbb{I}_{\{h(x)=0\}}. \tag{7}$$

All expectations and probabilities are taken w.r.t. the true data distribution $\mathcal{D}$. Throughout let $g(x, h)$ be the confidence score. For a target coverage $c \in (0, 1]$ denote by

$$t_c \quad \text{s.t.} \quad \Pr\big(g(X, h) \geq t_c\big) = c \tag{8}$$

the *population threshold*, and write the *accepted region* $A_c := \{x : g(x, h) \geq t_c\}$. The oracle that attains the perfect-ordering bound accepts $A_c^\star := \big\{x : \eta_h(x) \text{ is among the largest } c\text{-fraction}\big\}$.

Error terms. We isolate the following sources of error affecting selective prediction performance:

$$\varepsilon_{\text{Bayes}}(c) := \mathbb{E}\Big[1 - \max\{\eta(X), 1 - \eta(X)\} \;\Big|\; X \in A_c\Big], \tag{9}$$

$$\varepsilon_{\text{approx}}(c) := \mathbb{E}\Big[\big|\eta_h(X) - \eta(X)\big| \;\Big|\; X \in A_c\Big], \tag{10}$$

$$\varepsilon_{\text{rank}}(c) := \mathbb{E}\big[\eta_h(X) \mid X \in A_c^\star\big] - \mathbb{E}\big[\eta_h(X) \mid X \in A_c\big] \qquad (\geq 0), \tag{11}$$

$$\varepsilon_{\text{stat}}(c) := C\sqrt{\frac{\log(1/\delta)}{n}}, \tag{12}$$

where n is the evaluation-set size, $\delta \in (0, 1)$ a confidence parameter, and $C > 0$ an absolute constant. Intuitively, $\varepsilon_{\text{Bayes}}$ is the irreducible label noise inside the accepted region; $\varepsilon_{\text{approx}}$ measures how far h is from Bayes-optimal on the *selected* inputs; $\varepsilon_{\text{rank}}$ is a *ranking regret* measuring the accuracy loss due solely to picking the wrong c-fraction of samples; and $\varepsilon_{\text{stat}}$ captures the *sampling uncertainty* due to evaluating on a finite dataset. Note that we freeze the acceptance set A_c defined by the current scoring function and ask how much worse the learned classifier h is than the Bayes-optimal rule.

> **Remark** (Distance to a Perfect Ranker)**.** A natural way to gauge how far the learned acceptance rule is from the oracle is the *mass mis-ordered*
>
> $$D_{\text{rank}}(c) \;:=\; \Pr\big(X \in A_c^\star \setminus A_c\big) + \Pr\big(X \in A_c \setminus A_c^\star\big). \tag{13}$$
>
> It equals the total probability of examples that would have to be *swapped* between A_c and $A_c^\star$ to recover perfect ordering. Hence $D_{\text{rank}}(c) = 0$ iff $A_c = A_c^\star$, in which case $\varepsilon_{\text{rank}}(c)$ also vanishes.

> **Theorem 1** (Selective Gap Bound)**.** For a coverage level $c \in (0, 1]$ and a selective classifier (h, g) the population gap obeys
>
> $$\Delta(c) = \overline{\text{acc}}\big(a_{\text{full}}, c\big) - \text{acc}_c(h, g) \;\leq\; \varepsilon_{\text{Bayes}}(c) + \varepsilon_{\text{approx}}(c) + \varepsilon_{\text{rank}}(c). \tag{14}$$

Let $\widehat{\Delta}(c)$ be the empirical gap on n i.i.d. test points. Then, with probability at least $1-\delta$,

$$\widehat{\Delta}(c) \;\le\; \varepsilon_{\text{Bayes}}(c) + \varepsilon_{\text{approx}}(c) + \varepsilon_{\text{rank}}(c) + C\sqrt{\tfrac{\log(1/\delta)}{n}}. \tag{15}$$

Proof. Because $\mathrm{acc}_c(h,g) = \mathbb{E}[\eta_h(X) \mid A_c]$, the gap decomposes as

$$\Delta(c) = \underbrace{\mathbb{E}[\eta_h \mid A_c^\star] - \mathbb{E}[\eta_h \mid A_c]}_{\varepsilon_{\text{rank}}(c)} + \underbrace{\mathbb{E}[\eta_h - \mathbb{I}_{\{h=Y\}} \mid A_c]}_{\varepsilon_{\text{approx}}(c)} + \underbrace{\mathbb{E}[1-\max\{\eta, 1-\eta\} \mid A_c]}_{\varepsilon_{\text{Bayes}}(c)}.$$

This yields the population bound (14). For each expectation in the decomposition apply Hoeffding's inequality, a union bound over the three terms gives, with probability $1-\delta$, $|\widehat{\Delta}(c) - \Delta(c)| \le C\sqrt{\log(1/\delta)/n}$. Adding this deviation to (14) establishes (15).
See Appendix B.1 for an extended proof with detailed intermediate steps. □

A single design choice can shrink multiple error terms. We note that the individual error terms from the decomposition in Equation (15) can still interact with each other. For example, when the confidence score is the maximum softmax probability (MSP), a better approximation of the true conditional η not only lowers the *approximation* term $\varepsilon_{\text{approx}}(c)$ but also tends to align MSP more closely with η_h, thereby *indirectly reducing* the ranking error $\varepsilon_{\text{rank}}(c)$. Conversely, a non-monotone calibration head can reduce $\varepsilon_{\text{rank}}(c)$ without improving $\varepsilon_{\text{approx}}(c)$.

3.4 Calibration and Its (Limited) Effect on the Gap

As shown in Theorem 1, the selective classification gap includes a *ranking error* term $\varepsilon_{\text{rank}}(c)$, which captures misalignment between the confidence score and true correctness. Model calibration [Niculescu-Mizil and Caruana, 2005]—widely used to reduce over- or underconfidence—is often assumed to improve this alignment by transforming scores to better reflect correctness likelihood. Yet its effect on selective performance remains ambiguous and context-dependent. Prior work has reached conflicting conclusions: Zhu et al. [2022] argue that calibration may degrade abstention behavior, while Galil et al. [2023] find that temperature scaling can improve selective prediction in practice. We show that the impact on the gap depends critically on the *type* of calibration method used and its influence on the induced ranking. We begin by recalling the formal definition of calibration.

Definition 5 (Perfect Calibration)**.** For each input x let a model produce a predicted label $\hat{y}(x)$ and an associated confidence score $s(x) \in [0,1]$. We say the model is *perfectly calibrated* if

$$\Pr\big(Y = \hat{y}(X) \;\big|\; s(X) = t\big) \;=\; t \qquad \text{for every confidence level } t \in [0,1]. \tag{16}$$

Practical estimators approximate (16) via a post-hoc map ϕ such that $\tilde{s}(x) = \phi(s(x))$ approaches prefect calibration. *Expected Calibration Error (ECE)* [Naeini et al., 2015] quantifies this closeness:

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|I_b|}{n} \left| \frac{1}{|I_b|} \sum_{i \in I_b} \mathbb{I}\{\hat{y}(x_i) = y_i\} - \frac{1}{|I_b|} \sum_{i \in I_b} \tilde{s}(x_i) \right|, \tag{17}$$

where I_b is the set of indices in bin b, n is the total number of examples, and B is the number of bins.

Monotone score-level calibration leaves the gap intact. Isotonic regression [Zadrozny and Elkan, 2002] and histogram binning [Zadrozny and Elkan, 2001] fit a *monotone* $\phi\colon [0,1] \to [0,1]$ that preserves score ordering. Because monotone maps preserve ordering, the acceptance set $A_c = \{x : \tilde{s}(x) \ge \tau_c\}$ is identical to the one obtained from $s(x)$; hence the selective accuracy $\mathrm{acc}_c(h,g)$ and the gap $\Delta(c) = \overline{\mathrm{acc}}(a_{\text{full}}, c) - \mathrm{acc}_c(h,g)$ are *unchanged*. Monotone calibration thereby reduces the approximation error $\varepsilon_{\text{approx}}(c)$ in Section 3.3 but leaves the ranking error $\varepsilon_{\text{rank}}(c)$ untouched.

The effect of temperature scaling on the SC gap. Temperature scaling, the most widely used post-hoc calibration technique, divides every logit vector $z(x) \in \mathbb{R}^K$ by a scalar $T > 0$,

$$p_j^{(T)}(x) = \frac{\exp(z_j(x)/T)}{\sum_k \exp(z_k(x)/T)}. \tag{18}$$

While this operation leads to a *monotone* rescaling of the *logits*, it can lead to a *non-monotone* rescaling of the *softmax probabilities*. Since the softmax function is non-linear with respect to the temperature parameter, temperature scaling can therefore change the ranking of samples by confidence. This re-ranking can lead to small but empirically validated improvements in selective classification performance, as measured by metrics like AUROC [Galil et al., 2023, Cattelan and Silva, 2023]. However, the magnitude of this effect is inherently limited (see Appendix B.2 for an extended discussion). While temperature scaling can refine the ordering, it does not fundamentally alter the underlying quality of the model's uncertainty estimates.

Moving the gap requires non-monotone scoring. To reduce the selective classification gap $\Delta(c)$, it is not enough to calibrate scores post-hoc using monotone mappings. One must actively change the ranking of accepted examples to better reflect their true likelihood of correctness. Achieving this typically requires non-monotone scoring mechanisms that incorporate richer, instance-specific information—such as deep ensembles (`DE`), self-adaptive training (`SAT`), or learned correctness heads $g_\psi(x)$ that map hidden representations to confidence estimates. These approaches leverage model diversity, stochasticity, or internal feature structure to distinguish samples that would otherwise receive identical or wrongly ordered confidence values under standard softmax outputs.

Why binning and vector scaling should not be used. Histogram binning [Naeini et al., 2015] and vector/Dirichlet scaling [Kull et al., 2019]—while widely to improve calibration—are poorly suited for selective classification. Histogram binning quantizes scores into a small number of bins, mapping wide score intervals to the same value and destroying within-bin ordering, which leads to effectively random selection among tied examples. Vector and Dirichlet scaling are post-hoc calibration methods that generalize temperature scaling by learning class-specific transformations of logits—vector scaling applies a linear transformation, while Dirichlet scaling interprets the logits as parameters of a Dirichlet distribution to better model uncertainty. Recent work by Le-Coz et al. [2024] shows that histogram binning and vector/Dirichlet scaling consistently degrade AUROC in selective classification. These results underscore our central claim: improving calibration does not guarantee better ranking. Reducing the selective classification gap requires score functions that explicitly learn to separate easy from hard examples, not just to produce better-calibrated probabilities.

Loss prediction as a multicalibration litmus test. A complementary view on how calibration connects to ranking ability arises from the notion of *multicalibration* [Hébert-Johnson et al., 2018], which requires that a model's confidence be calibrated not only overall but also across many subgroups of inputs. Recent work by Gollakota et al. [2025] shows that achieving strong multicalibration is equivalent to learning an accurate predictor of one's own loss—that is, training an auxiliary model to estimate, for each input, the probability that the base predictor will be correct. Viewed this way, reliability becomes a self-forecasting problem: if a model (or an auxiliary head) can successfully predict its own 0–1 loss, then its confidence scores must already be well aligned with correctness, leaving little residual ranking error. We formalize this equivalence in Appendix E and show, both theoretically and empirically, that the degree to which a model's loss can be predicted corresponds directly to the magnitude of the ranking-error term $\varepsilon_{\text{rank}}(c)$. In short, when no auxiliary predictor can outperform the model's own confidence scores at identifying its mistakes, the model is effectively multicalibrated and near the oracle frontier; conversely, any nontrivial loss-prediction advantage exposes where—and by how much—its internal ranking deviates from perfect ordering.

> **Takeaway.** While post-hoc calibration with temperature scaling can provide modest improvements to ranking, it is not sufficient to close the SC gap. Substantially reducing the ranking error ($\varepsilon_{\text{rank}}$) requires more powerful scoring methods that actively re-rank examples based on richer information, such as feature-aware heads, ensembles, or non-monotone transformations.

3.5 Additional Practical Sources of Looseness

The decomposition in Theorem 1 captures the *intrinsic* sources of error—Bayes noise, approximation limits, ranking error, and sampling slack—forming a principled bound that holds even under perfect optimization, infinite data, and i.i.d. testing. In practical deployments, however, additional imperfections can inflate the empirical gap $\widehat{\Delta}(c)$. These stem from implementation details, scoring granularity, and distribution shift—not fundamental limits, but contingent slack terms reducible through better engineering. We summarize them below under a single *residual slack* term $\varepsilon_{\text{misc}}(c)$.

1. **Optimization error ε_{opt}.** In practice, gradient-based solvers rarely attain the empirical risk minimizer. If $L(\theta)$ denotes the end-to-end training objective—encompassing model architecture, loss (e.g. cross-entropy), and training data—and $\hat{\theta}$ its final iterate, then $\varepsilon_{\text{opt}} = L(\hat{\theta}) - \min_\theta L(\theta)$, which—via standard surrogate-to-0/1 calibration bounds—translates into a nonzero selective-accuracy loss that persists even under infinite data.
2. **Distribution shift $\varepsilon_{\text{shift}}(c)$.** When the test distribution p_{test} deviates from the training distribution p_{train}, both calibration and ranking typically degrade. In particular, for a hypothesis class $\mathcal{H}$, the gap due to shift can be bounded by an *Integral Probability Metric (IPM)* [Müller, 1997]:

$$\varepsilon_{\text{shift}}(c) \;\le\; \text{IPM}_{\mathcal{H}}\big(p_{\text{train}},\, p_{\text{test}}\big) := \sup_{f\in\mathcal{H}} \left|\mathbb{E}_{p_{\text{train}}}[f] - \mathbb{E}_{p_{\text{test}}}[f]\right|. \tag{19}$$

   Hence, larger shifts in distribution (relative to $\mathcal{H}$) lead to wider selective classification gaps.

Residual slack. The dominant practical sources of looseness are optimization error and distribution shift, summarized by $\varepsilon_{\text{misc}}(c) := \varepsilon_{\text{opt}} + \varepsilon_{\text{shift}}(c)$. These two terms capture the main drivers of residual deviation between the theoretical and empirical gaps. For completeness, we discuss additional minor contributors such as threshold-selection noise or score quantization in Appendix B.3. Together, these effects make the bound in Equation (20) *sufficient*, not merely necessary, for explaining all observed looseness in practical selective classifiers, yielding the streamlined high-probability bound.

$$\widehat{\Delta}(c) \;\le\; \underbrace{\varepsilon_{\text{Bayes}}(c) + \varepsilon_{\text{approx}}(c) + \varepsilon_{\text{rank}}(c) + \varepsilon_{\text{stat}}(c)}_{\text{intrinsic}} \;+\; \varepsilon_{\text{misc}}(c). \tag{20}$$

Takeaway. Only $\varepsilon_{\text{Bayes}}$ reflects irreducible uncertainty; the other intrinsic terms—$\varepsilon_{\text{approx}}$, $\varepsilon_{\text{rank}}$, and $\varepsilon_{\text{stat}}$—can be reduced with better models, calibration, and data. The *miscellaneous slack* $\varepsilon_{\text{misc}}$ highlights optimization and shift-robustness as levers for closing the gap to the oracle.

4 Empirical Results

Our experimental study is organized around three guiding questions that reflect the theoretical decomposition in Section 3. Unless otherwise specified, all results are averaged over 5 random seeds.

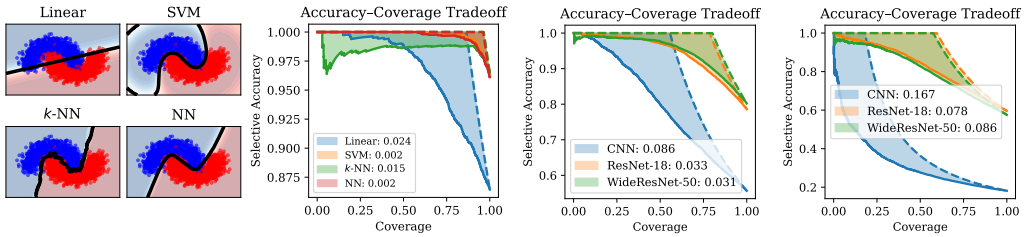
4.1 Q1: How do Bayes error and approximation error shape the gap?

Setup. We conduct both synthetic and real-world experiments. For our synthetic results, which give us precise control over the data generation process, we simulate two sources of intrinsic difficulty on the two-moons dataset: (i) **noise** $\sigma \in \{0.1,\, 0.33, 0.66,\, 1.5\}$ controls how much the two moons expand into each other; and (ii) **model capacity**, varied from logistic regression (low capacity) to a shallow MLP (high capacity). For our real-world experiments we tackle the analysis similarly: for (a) we evaluate a trained CIFAR-10 model on the CIFAR-10N/100N [Wei et al., 2022] datasets to assess which data points have large labeling disagreement; and for (b) we vary the model architecture across a simple CNN (details in Appendix D.3), a ResNet-18 [He et al., 2016], and a WideResNet-50 [Zagoruyko and Komodakis, 2016] on CIFAR-100 [Krizhevsky et al., 2009] and StanfordCars [Krause et al., 2013]. For each setting, we compute the Excess-AURC (E-AURC) [Geifman et al., 2019] by integrating the empirical gap $\widehat{\Delta}(c)$ across all coverage levels.

Findings. In terms of approximation error, Figure 2 demonstrates that limited model capacity leads to larger gaps, while more expressive models yield tighter alignment with the perfect-ordering bound. This suggests that approximation error is a key driver of looseness. In terms of Bayes error, Figure 3 shows that increasing label noise consistently lowers the accuracy–coverage curve, indicating that Bayes error introduces an irreducible component to the gap. These results validate the canonical bound (Equation (15)): large Bayes or approximation error can explain substantial looseness.

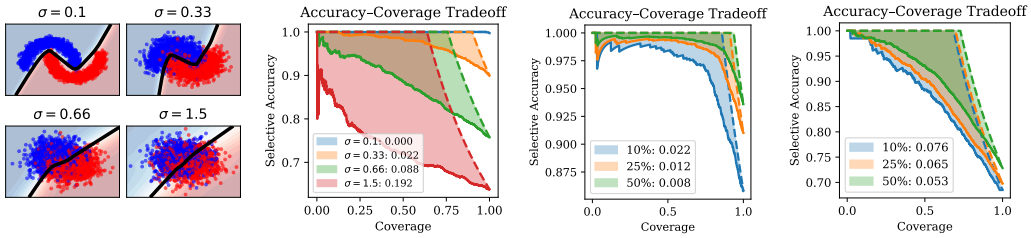
4.2 Q2: When—and what kind of—calibration helps?

Setup. We study the same three model classes as before on CIFAR-100: a lightweight CNN, a ResNet-18, and a WideResNet-50. On each backbone we evaluate the following confidence–scoring variants: (i) maximum softmax probability (`MSP`) [Hendrycks and Gimpel, 2017];



(a) Approximation error with two moons dataset. (b) CIFAR-100 (c) StanfordCars

Figure 2: **Experiments on approximation error**. We find that approximation error is a major contributor to the gap. (a) We show the two moons dataset fitted with models of different degrees of expressiveness as well as the corresponding accuracy-coverage tradeoffs. (b) + (c) Accuracy-coverage tradeoffs for various model architectures on CIFAR-100 and StanfordCars, respectively.



(a) Bayes error with two moons dataset. (b) CIFAR-10N (c) CIFAR-100N

Figure 3: **Experiments on Bayes error**. We find that irreducible noise significantly contributes to the gap. (a) We show the two moons dataset with varying degrees of noise $\sigma \in \{0.1, 0.33, 0.66, 1.5\}$ as well as the corresponding accuracy-coverage tradeoffs. (b) + (c) Accuracy-coverage tradeoffs for the 10% (blue), 25% (orange), and 50% (green) most noisy images in CIFAR-10N/100N, respectively.

(ii) a temperature-scaled softmax (monotone probability calibration, `TEMP`) [Guo et al., 2017]; (iii) self-adaptive training (`SAT`) [Huang et al., 2020], which implicitly calibrates by relabelling uncertain samples during training; and (iv) deep ensembles (`DE`) [Lakshminarayanan et al., 2017] of five independently initialised networks (non-monotone aggregation; improves ranking via variance). Our inclusion of the `MSP` baseline is motivated by the large-scale study of Jaeger et al. [2023], who find that `MSP`, while simple and easy to implement, is often hard-to-beat in practice. For each score we report (a) the weighted Expected Calibration Error (ECE); and (b) the Excess-AURC (E-AURC) [Geifman et al., 2019] metric measuring selective prediction performance.

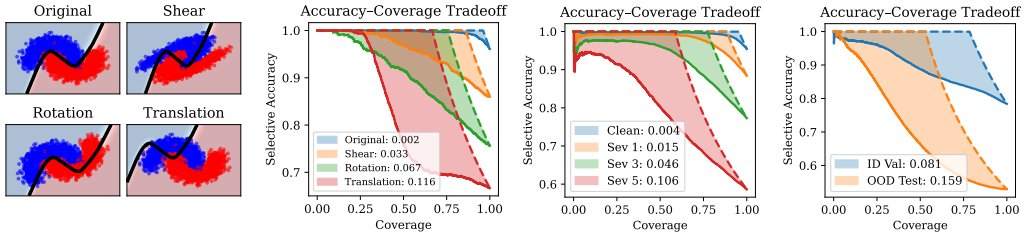
Findings. We summarize our findings in Table 1. While temperature scaling (`TEMP`) consistently improves ECE across model classes relative to `MSP`, it leaves the selective classification gap largely unchanged—highlighting the limitations of monotone calibration. In contrast, `SAT` slightly improves both ECE and gap by perturbing rankings through relabeling, while deep ensembles (`DE`) achieve the largest gap reductions by explicitly reordering predictions via averaging. These trends confirm that only methods capable of re-ranking—implicitly (`SAT`) or explicitly (`DE`)—can meaningfully improve selective performance. Consistent with this, we find that only `SAT` and `DE` models reliably predict their own loss, reinforcing their stronger alignment with correctness. See Appendix E.5 for details.

4.3 Q3: How does the gap evolve under distribution shift?

Setup. As in Q1, we explore this question using both synthetic and real-world distribution shifts. For synthetic experiments, we use the two moons dataset with three types of input shift: shear, rotation, and translation (details in Appendix D.4). For real data with synthetic corruptions, we use CIFAR-10C [Hendrycks and Dietterich, 2019], which applies algorithmic covariate corruptions to the CIFAR-10 test set across five severity levels (1–5). To evaluate under a real distribution shift, we also consider Camelyon17-WILDS [Koh et al., 2021]—a cancer detection dataset where test data is collected from a different hospital system than the training data.

Table 1: **Experiments on calibration across model classes on CIFAR-100**. Temperature scaling (TEMP) significantly improves ECE over the Maximum Softmax Probability (MSP) baseline but does not help to close the selective classification gap. Self-Adaptive Training (SAT) and Deep Ensembles (DE) improve calibration non-monotonically and also improve selective classification acceptance ordering through re-ranking. A corresponding plot is given in Figure 5; more datasets in Tables 2, 3.

	CNN				ResNet-18				WideResNet-50			
	MSP	TEMP	SAT	DE	MSP	TEMP	SAT	DE	MSP	TEMP	SAT	DE
E-AURC	0.086	0.085	0.081	0.065	0.033	0.033	0.028	0.026	0.031	0.032	0.028	0.026
ECE	0.142	0.008	0.116	0.019	0.052	0.048	0.026	0.034	0.066	0.050	0.046	0.030



(a) Distribution shifts with two moons dataset. (b) CIFAR-10C (c) Camelyon17-WILDS

Figure 4: **Experiments on distribution shifts**. We find that shifts can also significantly contribute to the gap. (a) Two moons under shear, rotation, and translation with corresponding accuracy–coverage curves. (b) CIFAR-10C across three distinct corruption severities. (c) Camelyon17 OOD shift.

Findings. Figure 4 shows a clear trend: as covariate shifts intensify, the accuracy–coverage curve moves farther below its oracle bound, indicating that abstention no longer isolates easy inputs. Selective classifiers thus become *over-confidently wrong*, echoing evidence that many uncertainty metrics deteriorate under shift or misspecification [Snoek et al., 2019]. As the gap grows with shift severity, deployments must pair selective prediction with robust ranking or shift-detection safeguards.

5 Conclusion

Building a truly performant selective classifier hinges on understanding and closing the gap between practical models and the oracle perfect-ordering bound. To answer *what it takes*, we introduce a coverage-uniform selective-classification gap and derive the first finite-sample decomposition that pinpoints exactly five limiting factors: three intrinsic sources—Bayes noise, approximation error, and ranking (calibration) error—and two contingent slack terms—sampling variability and implementation or distribution-shift artifacts. Our experiments show that each component can be individually measured and, importantly, directly improved: stronger model backbones reduce approximation error, non-monotone or feature-aware scoring shrinks ranking error, and shift-robust training with larger validation sets minimizes residual slack. Together, these insights provide a clear recipe for designing and evaluating high-performance selective classifiers.

Limitations and future work. While our decomposition cleanly bounds the selective-classification gap, its error budgets can *interact*—for example, increasing capacity often improves both approximation and ranking—which makes unique attribution challenging. Many *training-time calibration schemes* (e.g., SAT, mixup, focal loss) simultaneously affect ranking and full-coverage accuracy, confounding the separation of budgets. Our core experiments focus on *synthetic and vision benchmarks*; extending these insights to large-scale foundation models would be an important direction. We present a preliminary exploration on large language models in Appendix F.2. Finally, because our oracle bound and gap are defined for *0–1 loss*, adapting to *asymmetric or class-dependent cost functions*—often required in high-stakes decision-making—will require generalizing both the bound and its decomposition. Our finite-sample gap decomposition lays the groundwork for a more unified reliability framework; extending it to (i) settings where out-of-distribution inputs must be rejected and (ii) open-ended language generation constitutes a promising agenda for future work.

Acknowledgements

We acknowledge the following sponsors, who support our research with financial and in-kind contributions: Apple, CIFAR through the Canada CIFAR AI Chair, Meta, NSERC through the Discovery Grant and an Alliance Grant with ServiceNow and DRDC, the Ontario Early Researcher Award, the Schmidt Sciences foundation through the AI2050 Early Career Fellow program. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute. We thank Relu Patrascu and the computing team at the University of Toronto's Computer Science Department for administrating and procuring the compute infrastructure used for the experiments in this paper. We would also like to thank Andy Wei Liu, Anvith Thudi, David Glukhov, Vardan Papyan, and many others at the Vector Institute for discussions contributing to this paper.

References

A. N. Angelopoulos and S. Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *ArXiv preprint*, abs/2107.07511, 2021. URL https://arxiv.org/abs/2107.07511.

A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

C. M. Bishop. Pattern recognition and machine learning. *Springer*, 2:1122–1128, 2006.

M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9630–9640. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00951. URL https://doi.org/10.1109/ICCV48922.2021.00951.

L. F. P. Cattelan and D. Silva. How to fix a broken confidence estimator: Evaluating post-hoc methods for selective classification with deep neural networks. 2023.

T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020. URL http://proceedings.mlr.press/v119/chen20j.html.

C.-K. Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, (4):247–254, 1957.

P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv preprint*, abs/1803.05457, 2018. URL https://arxiv.org/abs/1803.05457.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

L. Coenen, A. K. A. Abdullah, and T. Guns. Probability of default estimation, with a reject option. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 439–448, 2020. doi: 10.1109/DSAA49011.2020.00058.

C. Corbière, N. Thome, A. Bar-Hen, M. Cord, and P. Pérez. Addressing failure prediction by learning model confidence. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2898–2909, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/757f843a169cc678064d9530d12a1881-Abstract.html.

E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation policies from data. *ArXiv preprint*, abs/1805.09501, 2018. URL https://arxiv.org/abs/1805.09501.

E. D. Cubuk, B. Zoph, J. Shlens, and Q. Le. Randaugment: Practical automated data augmentation with a reduced search space. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d85b63ef0ccb114d0a3bb7b7d808028f-Abstract.html.

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

A. Dietmüller, A. G. Alcoz, and L. Vanbever. Fitnets: An adaptive framework to learn accurate traffic distributions. *ArXiv preprint*, abs/2405.10931, 2024. URL https://arxiv.org/abs/2405.10931.

Y. Ding, J. Liu, J. Xiong, and Y. Shi. Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 4–5, 2020.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

R. El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

L. Feng, S. Shu, Z. Lin, F. Lv, L. Li, and B. An. Can cross entropy loss be robust to label noise? In C. Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 2206–2212. ijcai.org, 2020. doi: 10.24963/ijcai.2020/305. URL https://doi.org/10.24963/ijcai.2020/305.

I. Galil, M. Dabbah, and R. El-Yaniv. What can we learn from the selective prediction and uncertainty estimation performance of 523 imagenet classifiers? In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=p66AzKi6Xim.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17 (59):1–35, 2016.

Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4878–4887, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Y. Geifman and R. El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2151–2159. PMLR, 2019. URL http://proceedings.mlr.press/v97/geifman19a.html.

Y. Geifman, G. Uziel, and R. El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=SJfb5jCqKm.

Z. Ghodsi, S. K. S. Hari, I. Frosio, T. Tsai, A. Troccoli, S. W. Keckler, S. Garg, and A. Anandkumar. Generating and characterizing scenarios for safety testing of autonomous vehicles. *ArXiv preprint*, abs/2103.07403, 2021. URL https://arxiv.org/abs/2103.07403.

A. Gollakota, P. Gopalan, A. Karan, C. Peale, and U. Wieder. When does a predictor know its own loss? *ArXiv preprint*, abs/2502.20375, 2025. URL https://arxiv.org/abs/2502.20375.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

J. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent - A new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

H. Guan, Y. Zhang, H.-D. Cheng, and X. Tang. Bounded-abstaining classification for breast tumors in imbalanced ultrasound images. *International Journal of Applied Mathematics and Computer Science*, 30(2), 2020.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017. URL http://proceedings.mlr.press/v70/guo17a.html.

P. E. Hart, D. G. Stork, and R. Duda. *Pattern classification*. Wiley Hoboken, 2001.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Ú. Hébert-Johnson, M. P. Kim, O. Reingold, and G. N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In J. G. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1944–1953. PMLR, 2018. URL http://proceedings.mlr.press/v80/hebert-johnson18a.html.

D. Hendrycks and T. G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

D. Hendrycks, K. Lee, and M. Mazeika. Using pre-training can improve model robustness and uncertainty. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2712–2721. PMLR, 2019. URL http://proceedings.mlr.press/v97/hendrycks19a.html.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *ArXiv preprint*, abs/1503.02531, 2015. URL https://arxiv.org/abs/1503.02531.

G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer, 2016.

G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger. Snapshot ensembles: Train 1, get M for free. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=BJYwwY9ll.

J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In B. Schölkopf, J. C. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 601–608. MIT Press, 2006. URL https://proceedings.neurips.cc/paper/2006/hash/a2186aa7c086b46ad4e8bf81e2a3a19b-Abstract.html.

L. Huang, C. Zhang, and H. Zhang. Self-adaptive training: beyond empirical risk minimization. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e0ab531ec312161511493b002f9be2ee-Abstract.html.

P. F. Jaeger, C. T. Lüth, L. Klein, and T. J. Bungert. A call to reflect on evaluation practices for failure detection in image classification. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=YnkGMIh0gvX.

E. Jones, S. Sagawa, P. W. Koh, A. Kumar, and P. Liang. Selective classification can magnify disparities across groups. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=N0M_4BkQ05i.

S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, et al. Language models (mostly) know what they know. *ArXiv preprint*, abs/2207.05221, 2022. URL https://arxiv.org/abs/2207.05221.

P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. Earnshaw, I. S. Haque, S. M. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang. WILDS: A benchmark of in-the-wild distribution shifts. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 2021. URL http://proceedings.mlr.press/v139/koh21a.html.

R. Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

M. Kull, M. Perelló-Nieto, M. Kängsepp, T. de Menezes e Silva Filho, H. Song, and P. A. Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 12295–12305, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8ca01ea920679a0fe3728441494041b9-Abstract.html.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6402–6413, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

A. Le-Coz, S. Herbin, and F. Adjed. Confidence calibration of classifiers with many classes. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/8df80d7115a55eb81010c967a247b1ae-Abstract-Conference.html.

Y. Li, N. Wang, J. Shi, J. Liu, and X. Hou. Revisiting batch normalization for practical domain adaptation. *ArXiv preprint*, abs/1603.04779, 2016. URL https://arxiv.org/abs/1603.04779.

Z. Liu, Z. Wang, P. P. Liang, R. Salakhutdinov, L. Morency, and M. Ueda. Deep gamblers: Learning to abstain with portfolio theory. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10622–10632, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/0c4b1eeb45c90b52bfb9d07943d855ab-Abstract.html.

Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11966–11976. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01167. URL https://doi.org/10.1109/CVPR52688.2022.01167.

I. Loshchilov and F. Hutter. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

I. Loshchilov, F. Hutter, et al. Fixing weight decay regularization in adam. *ArXiv preprint*, abs/1711.05101, 2017. URL https://arxiv.org/abs/1711.05101.

P. Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

P. Micikevicius, S. Narang, J. Alben, G. F. Diamos, E. Elsen, D. García, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu. Mixed precision training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=r1gs9JgRZ.

B. Mucsányi, M. Kirchhof, and S. J. Oh. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/5afa9cb1e917b898ad418216dc726fbd-Abstract-Datasets_and_Benchmarks_Track.html.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443, 1997.

Z. Nado, N. Band, M. Collier, J. Djolonga, M. W. Dusenberry, S. Farquhar, Q. Feng, A. Filos, M. Havasi, R. Jenatton, et al. Uncertainty baselines: Benchmarks for uncertainty & robustness in deep learning. *ArXiv preprint*, abs/2106.04015, 2021. URL https://arxiv.org/abs/2106.04015.

M. P. Naeini, G. F. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In B. Bonet and S. Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2901–2907. AAAI Press, 2015. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9667.

A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In L. D. Raedt and S. Wrobel, editors, *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, volume 119 of *ACM International Conference Proceeding Series*, pages 625–632. ACM, 2005. doi: 10.1145/1102351.1102430. URL https://doi.org/10.1145/1102351.1102430.

C. Northcutt, L. Jiang, and I. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky. Human uncertainty makes classification more robust. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9616–9625. IEEE, 2019. doi: 10.1109/ICCV.2019.00971. URL https://doi.org/10.1109/ICCV.2019.00971.

A. Pugnana and S. Ruggieri. Auc-based selective classification. In F. J. R. Ruiz, J. G. Dy, and J. van de Meent, editors, *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, volume 206 of *Proceedings of Machine Learning Research*, pages 2494–2514. PMLR, 2023. URL https://proceedings.mlr.press/v206/pugnana23a.html.

S. Rabanser, S. Günnemann, and Z. C. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1394–1406, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/846c260d715e5b854ffad5f70a516c88-Abstract.html.

S. Rabanser, A. Thudi, K. Hamidieh, A. Dziedzic, and N. Papernot. Selective classification via neural network training dynamics. *ArXiv preprint*, abs/2205.13532, 2022. URL https://arxiv.org/abs/2205.13532.

S. Rabanser, A. Thudi, A. G. Thakurta, K. Dvijotham, and N. Papernot. Training private models that know what they don't know. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a8526465a91166fbb90aaa8452b21eda-Abstract-Conference.html.

S. Rabanser, A. S. Shamsabadi, O. Franzese, X. Wang, A. Weller, and N. Papernot. Confidential guardian: Cryptographically prohibiting the abuse of model abstention. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

R. Salakhutdinov. Deep learning. In S. A. Macskassy, C. Perlich, J. Leskovec, W. Wang, and R. Ghani, editors, *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, page 1973. ACM, 2014. doi: 10.1145/2623330.2630809. URL https://doi.org/10.1145/2623330.2630809.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

L. N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017.

L. N. Smith and N. Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE, 2019.

J. Snoek, Y. Ovadia, E. Fertig, B. Lakshminarayanan, S. Nowozin, D. Sculley, J. V. Dillon, J. Ren, and Z. Nado. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13969–13980, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958, 2014.

G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al. Gemma 3 technical report. *ArXiv preprint*, abs/2503.19786, 2025. URL https://arxiv.org/abs/2503.19786.

R. J. Tibshirani and B. Efron. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57(1):1–436, 1993.

J. Traub, T. J. Bungert, C. T. Lüth, M. Baumgartner, K. H. Maier-Hein, L. Maier-Hein, and P. F. Jaeger. Overcoming common flaws in the evaluation of selective classification systems. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/047c84ec50bd8ea29349b996fc64af4b-Abstract-Conference.html.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer, 2005.

D. Wang, E. Shelhamer, S. Liu, B. A. Olshausen, and T. Darrell. Tent: Fully test-time adaptation by entropy minimization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, and W. Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html.

J. Wei, Z. Zhu, H. Cheng, T. Liu, G. Niu, and Y. Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=TBWA6PLJZQm.

Y. Wiener and R. El-Yaniv. Agnostic selective classification. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*,

pages 1665–1673, 2011. URL https://proceedings.neurips.cc/paper/2011/hash/4b6538a44a1dfdc2b83477cd76dee98e-Abstract.html.

S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5987–5995. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.634. URL https://doi.org/10.1109/CVPR.2017.634.

S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6022–6031. IEEE, 2019. doi: 10.1109/ICCV.2019.00612. URL https://doi.org/10.1109/ICCV.2019.00612.

B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In C. E. Brodley and A. P. Danyluk, editors, *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 609–616. Morgan Kaufmann, 2001.

B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.

S. Zagoruyko and N. Komodakis. Wide residual networks. In R. C. Wilson, E. R. Hancock, and W. A. P. Smith, editors, *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*. BMVA Press, 2016. URL http://www.bmva.org/bmvc/2016/papers/paper087/index.html.

H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

F. Zhu, Z. Cheng, X.-Y. Zhang, and C.-L. Liu. Rethinking confidence calibration for failure prediction. In *European conference on computer vision*, pages 518–536. Springer, 2022.

A Broader Impact

This work introduces a decomposition of the selective-classification gap into measurable components—Bayes noise, approximation error, ranking error, statistical noise, and deployment slack—offering practical guidance for improving abstaining classifiers. By diagnosing which source dominates in a given setting, our method supports more targeted model design and evaluation.

Positive implications. Our decomposition improves transparency and supports safer deployment in high-stakes domains by helping practitioners understand whether their model underperforms due to ranking, capacity, or robustness. Because each gap component is explicitly quantified, our approach can serve as a tool for model debugging, monitoring, and fairer benchmarking.

Potential risks. Selective classifiers may disproportionately defer on certain groups, amplifying disparities—a risk previously observed by Jones et al. [2021]. Additionally, institutions may exploit uncertainty estimates to justify *strategic abstention*—deliberately deferring on individuals they prefer not to serve [Rabanser et al., 2025]. While our framework identifies which part of the gap drives poor performance, it does not control how deferred inputs are handled.

Mitigations. We recommend reporting gap components disaggregated by sensitive attributes, auditing scoring functions for spurious correlations, and documenting fallback policies. These steps are essential to ensure that abstention mechanisms improve reliability without undermining fairness.

Outlook. We hope this work encourages more precise evaluations of selective classifiers, shifting focus from aggregate calibration to interpretable, component-wise gap analysis that can inform both technical improvements and policy safeguards.

B Methods Extension

B.1 Detailed Proof of Theorem 1

We restate the theorem for convenience.

Theorem 2 (Selective classification Gap; detailed)**.** Fix a coverage level $c \in (0, 1]$, a score function $g(\cdot, h)$, and its associated population threshold t_c satisfying $\Pr\big(g(X, h) \geq t_c\big) = c$. Define the accepted region $A_c := \{x : g(x, h) \geq t_c\}$ and the oracle region $A_c^\star := \{x : \eta_h(x)$ is among the largest c-fraction$\}$. With the error terms

$$\varepsilon_{\text{Bayes}}(c) = \mathbb{E}\left[1 - \max\{\eta(X), 1 - \eta(X)\} \mid X \in A_c\right], \tag{21}$$

$$\varepsilon_{\text{approx}}(c) = \mathbb{E}\left[|\eta_h(X) - \eta(X)| \mid X \in A_c\right], \tag{22}$$

$$\varepsilon_{\text{rank}}(c) = \mathbb{E}\left[\eta_h(X) \mid X \in A_c^\star\right] - \mathbb{E}\left[\eta_h(X) \mid X \in A_c\right] \quad (\geq 0), \tag{23}$$

the population gap satisfies

$$\Delta(c) = \overline{\text{acc}}\big(a_{\text{full}}, c\big) - \text{acc}_c(h, g) \;\leq\; \varepsilon_{\text{Bayes}}(c) + \varepsilon_{\text{approx}}(c) + \varepsilon_{\text{rank}}(c). \tag{24}$$

Moreover, let $\widehat{\Delta}(c)$ be the empirical gap computed on n independent test samples. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\widehat{\Delta}(c) \;\leq\; \varepsilon_{\text{Bayes}}(c) + \varepsilon_{\text{approx}}(c) + \varepsilon_{\text{rank}}(c) + C\sqrt{\frac{\log(1/\delta)}{n}}, \tag{25}$$

where $C > 0$ is an absolute constant.

Proof. We split the argument into four self-contained steps.

Step 0. Oracle upper bound revisited. For completeness we justify the piecewise form of $\overline{\text{acc}}\big(a_{\text{full}}, c\big)$ in Definition 3. Because $a_{\text{full}} = \Pr(h(X) = Y) = \mathbb{E}[\eta_h(X)]$, the set $\{x : \eta_h(x) = 1\}$ has probability mass at least a_{full}. Hence an oracle that retains the

highest-confidence points achieves perfect accuracy for all coverages $c \le a_{\text{full}}$. For $c > a_{\text{full}}$, the best it can do is include *only* those perfect points plus a $(c - a_{\text{full}})$-fraction of the remaining examples, which contribute at worst zero accuracy. Therefore

$$\overline{\text{acc}}(a_{\text{full}}, c) = \frac{a_{\text{full}}}{c}, \qquad a_{\text{full}} < c < 1. \tag{26}$$

Step 1. Algebraic decomposition of the gap. Recall that $\text{acc}_c(h, g) = \mathbb{E}[\eta_h(X) \mid X \in A_c]$. We repeatedly add and subtract the same quantity:

$$\begin{aligned}
\Delta(c) &:= \overline{\text{acc}}(a_{\text{full}}, c) - \text{acc}_c(h, g) \\
&= \overline{\text{acc}}(a_{\text{full}}, c) - \mathbb{E}[\eta_h \mid A_c^\star] \;+\; \mathbb{E}[\eta_h \mid A_c^\star] - \mathbb{E}[\eta_h \mid A_c] \\
&\le \mathbb{E}[\eta_h \mid A_c^\star] - \mathbb{E}[\eta_h \mid A_c] && \text{(rank)} \\
&\quad + \mathbb{E}[\eta_h - \mathbb{I}_{\{h=Y\}} \mid A_c] && \text{(approx+Bayes)} \\
&= \varepsilon_{\text{rank}}(c) + \varepsilon_{\text{approx}}(c) + \varepsilon_{\text{Bayes}}(c). && (27)
\end{aligned}$$

Explanation of the two labelled inequalities.

1. **(rank)** isolates the ranking error, $\varepsilon_{\text{rank}}(c) := \mathbb{E}[\eta_h \mid A_c^\star] - \mathbb{E}[\eta_h \mid A_c]$. The inequality holds because the remaining term from the previous line, $\overline{\text{acc}}(a_{\text{full}}, c) - \mathbb{E}[\eta_h \mid A_c^\star]$, is a non-negative quantity that is bounded by the error sources introduced next.

2. **(approx+Bayes)** adds and subtracts $\eta(X)$ inside the expectation, then splits the absolute value:
$$\eta_h - I_{\{h=Y\}} = (\eta_h - \eta) + (\eta - I_{\{h=Y\}}). \tag{28}$$
The second summand satisfies the deterministic bound $|\eta(X) - I_{\{h=Y\}}| = \max\{\eta, 1-\eta\} - I_{\{h=Y\}} \le 1 - \max\{\eta, 1-\eta\}$, yielding exactly $\varepsilon_{\text{Bayes}}(c)$. The first summand contributes $\varepsilon_{\text{approx}}(c)$.

Step 2. Non-negativity of $\varepsilon_{\text{rank}}(c)$. Because $\eta_h(X) \in [0, 1]$ and $A_c^\star$ contains the c-fraction of points with the largest η_h-values, $\mathbb{E}[\eta_h \mid A_c^\star] \ge \mathbb{E}[\eta_h \mid A_c]$, hence $\varepsilon_{\text{rank}}(c) \ge 0$ as stated.

Step 3. Finite-sample deviation. Let $\widehat{\mu}$ be any empirical average of a $[0, 1]$-valued random variable with expectation μ. Hoeffding's inequality gives $\Pr(|\widehat{\mu} - \mu| > \epsilon) \le 2e^{-2n\epsilon^2}$. Apply this bound separately to the three empirical estimates that constitute $\widehat{\Delta}(c)$, and take a union bound with $\epsilon = \sqrt{\frac{\log(6/\delta)}{2n}}$. This yields, with probability at least $1 - \delta$, $|\widehat{\Delta}(c) - \Delta(c)| \le C\sqrt{\log(1/\delta)/n}$ for an absolute constant C. Combining with (24) proves (25).

Step 4. Connection to ranking distance. Define the mass of mis-ordered points $D_{\text{rank}}(c) := \Pr\big(X \in A_c^\star \setminus A_c\big) + \Pr\big(X \in A_c \setminus A_c^\star\big)$. Because $\eta_h \in [0, 1]$,

$$\begin{aligned}
\varepsilon_{\text{rank}}(c) &= \mathbb{E}[\eta_h \mid A_c^\star] - \mathbb{E}[\eta_h \mid A_c] && (29) \\
&\le \big\|\eta_h\big\|_\infty D_{\text{rank}}(c) && (30) \\
&\le D_{\text{rank}}(c). && (31)
\end{aligned}$$

Hence $\varepsilon_{\text{rank}}(c) = 0$ if and only if $A_c = A_c^\star$.

This completes the proof. □

Multiclass remark. For $K > 2$ labels, define $\eta(x) = \big(\Pr(Y = 1 \mid x), \ldots, \Pr(Y = K \mid x)\big)$ and its complement confidence $\eta^{\max}(x) = \max_k \eta_k(x)$. Then the inequality $|\eta^{\max} - I_{\{h=Y\}}| \le 1 - \eta^{\max}$ replaces the binary bound above, and the rest of the argument goes through verbatim. The approximation term becomes $\mathbb{E}[\|\eta_h - \eta\|_1 \mid A_c]$; all other quantities are unchanged.

B.2 When Can Temperature Scaling Re-rank Confidence Scores?

Temperature scaling multiplies every logit by the same factor $1/T$ ($T > 0$) before the softmax,

$$p_j^{(T)}(x) = \frac{\exp(z_j(x)/T)}{\sum_k \exp(z_k(x)/T)}. \tag{32}$$

Although the predicted label $\arg\max_j z_j(x)$ is invariant to T, the *confidence score* $s_T(x) = \max_j p_j^{(T)}(x)$ can change its *cross-sample* ordering.

General form. Let $j_\star = \arg\max_j z_j(x)$ and $r_j(x) = \exp\big(z_j(x) - z_{j_\star}(x)\big)$ ($j \neq j_\star$). Then

$$s_T(x) = \frac{1}{1 + \sum_{j \neq j_\star} r_j(x)^{1/T}}. \tag{33}$$

For binary classification, the sum has a single term and (33) collapses to the familiar logistic form $s_T(x) = 1/(1 + e^{-\Delta/T})$ with $\Delta = z_{j_\star} - z_{3-j_\star}$.

Two-sample condition. For two inputs x_1, x_2 let $S_i(T) = \sum_{j \neq j_\star^{(i)}} r_{ij}^{1/T}$. Because each $r_{ij} \leq 1$, every $r_{ij}^{1/T}$ is monotone non-decreasing in T (strictly increasing unless there is a tie), and the ordering $s_T(x_1) > s_T(x_2)$ can change exactly at those temperatures T where $S_1(T) = S_2(T)$.

Illustrative example ($K = 3$).

$$z^{(1)} = (-2, -3, -3), \quad z^{(2)} = (0, -0.1, -3). \tag{34}$$

At $T = 1$ one finds $s_1(x_1) = 0.576 > 0.512 = s_1(x_2)$, while at $T = 3$ we see that $s_3(x_1) = 0.411 < 0.428 = s_3(x_2)$, so temperature scaling would now accept x_2 before x_1.

How likely is a swap? Equation (33) shows that a swap requires the one-dimensional curves $S_1(T)$ and $S_2(T)$ to intersect. Since the curves are continuous and monotone, the intersection occurs— if at all—at isolated temperatures and only when the competing logit patterns are finely tuned.

Practical implication. Temperature scaling can *in principle* tighten the selective-classification gap, but only for the vanishingly small subset of inputs whose non-maximum logits happen to satisfy $S_1(T^\star) = S_2(T^\star)$. To obtain a meaningful re-ordering one must therefore adopt *non-monotone* calibration strategies.

B.3 Additional Contingent Slack

In the main text (Sec. 3.5) we folded all implementation-level imperfections into a single residual term $\varepsilon_{\text{misc}}(c)$, retaining only optimization error and distribution shift explicitly. Here we list two further slack terms omitted there:

3. **Threshold-selection noise $\varepsilon_{\text{thr}}(c)$.**
   When the coverage threshold $\hat{t}_c$ is chosen on a validation set of size m, the realized coverage deviates from the target c by
   $$O\big(\sqrt{c(1-c)/m}\big), \tag{35}$$
   inducing a corresponding vertical shift in selective accuracy.
4. **Tie-breaking / score quantization $\varepsilon_{\text{tie}}(c)$.**
   Discrete confidence values (e.g. low-precision logits) create equivalence classes of samples with identical scores. If κ denotes the maximum number of tied samples at any score level, then
   $$\varepsilon_{\text{tie}}(c) \;\leq\; \frac{\kappa}{n}, \tag{36}$$
   where n is the size of the evaluation set.

Residual slack revisited. Together with optimization error ε_{opt} and shift $\varepsilon_{\text{shift}}(c)$, these yield

$$\varepsilon_{\text{misc}}(c) = \varepsilon_{\text{opt}} + \varepsilon_{\text{shift}}(c) + \varepsilon_{\text{thr}}(c) + \varepsilon_{\text{tie}}(c). \tag{37}$$

C Practitioner Checklist for Tightening the Selective-Classification Gap

Below is an expanded, actionable checklist to help practitioners systematically tackle each component of the selective-classification gap. For each item, we list concrete steps, recommended tools, and pointers to reduce the corresponding error term.

- $\varepsilon_{\text{approx}}$ **— Shrink Approximation Error**
  - *Model capacity:* Upgrade to deeper or wider architectures like ResNeXt, ViT, or ConvNeXt to better approximate complex functions and reduce base error [Xie et al., 2017, Dosovitskiy et al., 2021, Liu et al., 2022, Kadavath et al., 2022].
  - *Pre-training:* Initializing with rich features from self-supervised methods (SimCLR, BYOL) or foundation models (CLIP, DINO) can improve out-of-the-box performance, convergence, and uncertainty scores [Chen et al., 2020, Grill et al., 2020, Radford et al., 2021, Caron et al., 2021, Hendrycks et al., 2019]. However, pre-training can also sometimes negatively affect selective classification performance [Galil et al., 2023].
  - *Distillation:* Use teacher–student training with logit matching or feature hints to inherit accuracy from a larger model at lower cost [Galil et al., 2023, Hinton et al., 2015, Dietmüller et al., 2024].
  - *Data augmentation:* Augmentations can often improve generalization with policy-based (AutoAugment, RandAugment) or mixing-based (MixUp, CutMix) augmentations to regularize the learner [Cubuk et al., 2018, 2020, Zhang et al., 2018, Yun et al., 2019]. However, strong augmentations may also degrade selective classification performance for certain minority classes [Jones et al., 2021].
- $\varepsilon_{\text{rank}}$ **— Improve Ranking Calibration:**
  - *Feature-aware scoring:* Train auxiliary heads like ConfidNet to learn correctness scores using both logits and input features [Corbière et al., 2019], often improving uncertainty estimates. Self-Adaptive Training (SAT) further enhances this by encouraging internal representations to separate correct and incorrect predictions through contrastive regularization or supervised signals [Huang et al., 2020].
  - *Deep ensembles:* Use the disagreement or predictive entropy across multiple independently trained models to estimate uncertainty [Lakshminarayanan et al., 2017].
  - *Conformal methods:* Generate conformal p-values or risk-controlled selection sets that respect desired coverage guarantees [Vovk et al., 2005, Angelopoulos et al., 2024].
  - *Use caution with vector/Dirichlet scaling:* While previous work has shown that vector, matrix, or Dirichlet transformations can be beneficial to reshape confidence distributions [Guo et al., 2017, Kull et al., 2019], Le-Coz et al. [2024] shows that these techniques can harm ranking under a large number of classes.
- ε_{opt} **— Reduce Optimization Error:**
  - *Convergence diagnostics:* Track training/validation loss curves to detect underfitting and determine optimal stopping points [Salakhutdinov, 2014].
  - *Learning-rate schedules:* Employ dynamic LR strategies like cosine decay, OneCycle, or CLR to reach better optima more consistently [Loshchilov and Hutter, 2017, Smith and Topin, 2019, Smith, 2017].
  - *Early stopping / checkpoints:* Save and average late-stage checkpoints or use snapshot ensembling to smooth optimization variance [Huang et al., 2017, Lakshminarayanan et al., 2017, Rabanser et al., 2022].
  - *Regularization:* Use dropout, weight decay, or stochastic depth to prevent overfitting and stabilize training [Srivastava et al., 2014, Huang et al., 2016, Loshchilov et al., 2017].
- $\varepsilon_{\text{Bayes}}$ **— Quantify Irreducible Noise:**
  - *Repeated labels:* Collect multiple annotations (e.g., CIFAR-10H) to estimate human-level disagreement and the Bayes error floor [Peterson et al., 2019, Wei et al., 2022].
  - *Noise-robust training:* Mitigate label noise using bootstrapped or Taylor-truncated loss functions that temper reliance on hard labels [Reed et al., 2014, Feng et al., 2020].
  - *Dataset curation:* Apply confident learning to flag likely label errors or use active learning for data relabeling [Northcutt et al., 2021].

- $\varepsilon_{\text{stat}}$ **— Control Statistical Slack:**
  - *Validation set size:* Use a sufficiently large holdout set to estimate thresholds and calibrate uncertainty reliably [Hart et al., 2001].
  - *Confidence intervals:* Use DKW or Clopper–Pearson bounds to set conservative thresholds with statistical guarantees on coverage [Massart, 1990, Clopper and Pearson, 1934].
  - *Cross-validation:* Average selection thresholds over folds to reduce their variance and avoid overfitting to a single validation set [Kohavi et al., 1995].
- $\varepsilon_{\text{shift}}$ **— Mitigate Distribution Shift:**
  - *Shift detection:* Detect covariate shift via statistical two-sample tests such as MMD or KL divergence between feature distributions [Gretton et al., 2012, Rabanser et al., 2019].
  - *Importance weighting:* Correct mismatched data distributions with density ratio weighting, e.g., using kernel mean matching [Huang et al., 2006].
  - *Domain adaptation:* Finetune with in-domain examples or use unsupervised techniques like AdaBN or domain-adversarial training (DANN) [Ganin et al., 2016, Li et al., 2016].
  - *Test-time adaptation:* Adapt models at inference using entropy minimization (Tent) or batch norm recalibration to restore accuracy under shift [Nado et al., 2021, Wang et al., 2021].
- ε_{thr} **— Threshold–Selection Noise:**
  - *Bootstrap resampling:* Estimate variability in the selection threshold τ_c by computing its standard error across bootstrap samples [Tibshirani and Efron, 1993].
  - *Smooth thresholds:* Interpolate between adjacent scores or accept a random subset at the threshold to reduce coverage discontinuities [Angelopoulos and Bates, 2021].
- ε_{tie} **— Tie-Breaking & Score Quantization:**
  - *Higher precision:* Use higher float precision (e.g., FP32 or FP64) or more logits bits to distinguish close scores and avoid ties [Micikevicius et al., 2018].
  - *Dithering:* Add tiny random noise to scores before thresholding to stochastically resolve ties and reduce instability.
  - *Refrain from binning:* Histogram binning (HQ) or Bayesian Binning into Quantiles (BBQ) often improve calibration but not selective classification performance [Naeini et al., 2015, Le-Coz et al., 2024].

Putting it all together. After addressing each bullet above, recompute your selective accuracy–coverage curve and compare to the oracle bound (Def. 3). Iterating over these steps will systematically shrink $\widehat{\Delta}(c)$ toward its irreducible floor.

D Experimental Details

D.1 Computational Resources

Our experiments were conducted on a mix of GPU-equipped compute nodes with varying hardware configurations. Some machines are equipped with Intel Xeon Silver CPUs (10 cores, 20 threads) and 128GB of RAM, each hosting 4× NVIDIA GeForce RTX 2080 Ti GPUs with 11GB VRAM. Others feature AMD EPYC 7643 processors (48 cores, 96 threads), 512GB of RAM, and 4× NVIDIA A100 GPUs, each with 80GB VRAM.

D.2 Hyper-Parameters

We follow standard literature-recommended training settings across all datasets. For each architecture–dataset pair, we use a fixed learning rate, weight decay, and batch size as detailed below:

- **SimpleCNN:**
  - Learning rate: 0.01
  - Weight decay: 1×10^{-4}
  - Batch size: 128
- **ResNet-18:**

  - Learning rate: 0.1 for CIFAR datasets; 0.01 for Stanford Cars, Camelyon17
  - Weight decay: 5×10^{-4}
  - Batch size: 128
- **WideResNet-50-2:**
  - Same settings as ResNet-18
- **Epochs:**
  - 200 epochs for all datasets except Camelyon17, which uses 10
- **Optimization:** SGD with momentum 0.9, Nesterov enabled, and a cosine annealing learning rate schedule.
- **Selective prediction methods:**
  - `MSP`: Standard cross-entropy training
  - `SAT`: Cross-entropy pretraining for half of training epochs, followed by Self-Adaptive Training (momentum 0.9) with an extra abstain class

All experiments use fixed random seeds for reproducibility and standard data augmentation per dataset (random crops, flips, normalization).

D.3 SimpleCNN Architecture

The SimpleCNN model is a compact convolutional neural network used for experiments on lower-resolution image datasets. The architecture is defined by the following sequence of layers:

- A 3×3 convolution with 32 filters and padding 1, followed by ReLU and 2×2 max-pooling.
- A second 3×3 convolution with 64 filters and padding 1, followed by ReLU and 2×2 max-pooling.
- A flattening layer, followed by a fully connected layer with 128 hidden units and ReLU activation.
- A final fully connected layer projecting to the number of output classes.

Let $s = \texttt{input\_size}//4$ denote the spatial resolution after two 2×2 pooling layers. Then, the full model is:

$$\begin{aligned}\texttt{SimpleCNN}(x) = {} & \texttt{Linear}\big(128 \to \texttt{num\_classes}\big) \circ \texttt{ReLU} \circ \\ & \texttt{Linear}\big(64 \cdot s^2 \to 128\big) \circ \texttt{Flatten} \circ \\ & \texttt{MaxPool2d} \circ \texttt{ReLU} \circ \texttt{Conv2d}(32 \to 64) \circ \\ & \texttt{MaxPool2d} \circ \texttt{ReLU} \circ \texttt{Conv2d}(3 \to 32)(x)\end{aligned}$$

The number of output classes is set as follows:

$$\texttt{num\_classes} = \begin{cases} 10 & \text{for CIFAR-10,} \\ 100 & \text{for CIFAR-100,} \\ 196 & \text{for Stanford Cars,} \\ 2 & \text{for Camelyon17,} \end{cases} \quad \text{with an optional extra class if } \texttt{extra\_class} \text{ is True.}$$

The input size is dataset-dependent and set to:

$$\texttt{input\_size} = \begin{cases} 32 & \text{for CIFAR-10 and CIFAR-100,} \\ 224 & \text{for Stanford Cars, Camelyon17.} \end{cases}$$

The model structure is summarized below:

```
SimpleCNN(
  (net): Sequential(
    (0): Conv2d(3, 32, kernel_size=(3, 3), stride=(1, 1), padding=1)
    (1): ReLU()
```

```
    (2): MaxPool2d(kernel_size=2, stride=2)
    (3): Conv2d(32, 64, kernel_size=(3, 3), stride=(1, 1), padding=1)
    (4): ReLU()
    (5): MaxPool2d(kernel_size=2, stride=2)
    (6): Flatten(start_dim=1)
    (7): Linear(in_features=4096, out_features=128)
    (8): ReLU()
    (9): Linear(in_features=128, out_features=num_classes)
  )
)
```

D.4 Synthetic Distribution Shifts on Two Moons

To evaluate robustness under controlled covariate shifts, we apply a series of synthetic affine transformations to the test set of the standard two moons dataset. Each transformation simulates a distinct type of distribution shift:

- **Original:** No transformation; the unperturbed test set.
- **Shear:** A shear transformation along the x-axis defined by:

$$\text{Shear matrix} \quad S = \begin{bmatrix} 1 & 1.25 \\ 0 & 1 \end{bmatrix}, \quad \text{so that} \quad x' = Sx = \begin{bmatrix} x + 1.25y \\ y \end{bmatrix}. \tag{38}$$

- **Rotation:** A rotation by 30 degrees counterclockwise, using:

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \quad \theta = \frac{\pi}{6}. \tag{39}$$

- **Translation:** A shift of the input space by a fixed vector:

$$x' = x + t, \quad \text{where} \quad t = \begin{bmatrix} 1.0 \\ -0.5 \end{bmatrix}. \tag{40}$$

Each transformation is applied to the test data matrix X_{test} via matrix multiplication or translation, yielding the following test sets:

$$\begin{aligned} \text{Original:} \quad & X_{\text{test}} \\ \text{Shear:} \quad & X_{\text{test}} \cdot S^\top \\ \text{Rotation:} \quad & X_{\text{test}} \cdot R^\top \\ \text{Translation:} \quad & X_{\text{test}} + t \end{aligned} \tag{41}$$

These transformations create meaningful distribution shifts while preserving label semantics, enabling precise evaluations of model robustness under shift.

D.5 CIFAR-10C Severity Levels

For the CIFAR-10C severity levels (1–5), we aggregate all 15 corruption types at a given severity to form a single validation set. For severity level l, we collect all corruptions labeled as severity l across the following categories:

- **Noise:** `gaussian_noise`, `shot_noise`, `impulse_noise`
- **Blur:** `defocus_blur`, `glass_blur`, `motion_blur`, `zoom_blur`
- **Weather:** `snow`, `frost`, `fog`, `brightness`
- **Digital:** `contrast`, `elastic_transform`, `pixelate`, `jpeg_compression`

This results in a single validation set per severity level l, where each image is sampled from one of these 15 corruptions applied at the specified severity.

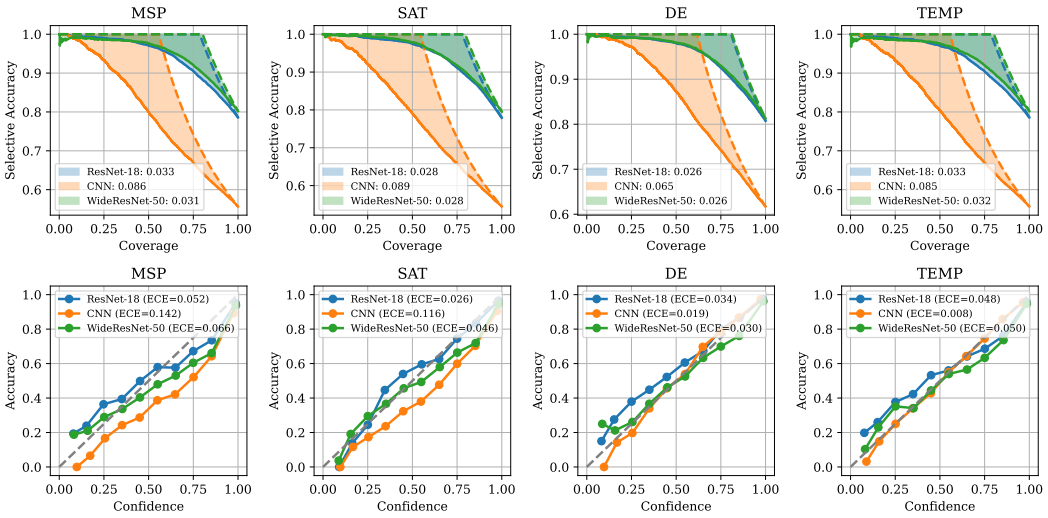


Figure 5: **Comparison between gap and calibration on CIFAR-100.** *Top*: selective accuracy curves across four training methods and three architectures. *Bottom*: corresponding reliability diagrams (ECE shown in parentheses). Temperature scaling (`TEMP`) consistently improves calibration but does not reduce the gap. By contrast, `SAT` and `DE` reduce the gap more effectively—especially for larger models—by improving the ranking.

E Loss Prediction, Multicalibration, and Ranking Error

This appendix offers an alternative perspective on the ranking error term $\varepsilon_{\text{rank}}(c)$ by framing it as a challenge of per-example loss prediction. Instead of building directly on the calibration discussion in Section 3.4, we show how the ability to forecast one's own 0–1 loss tightly controls the selective-classification gap. We formalize this connection through the recent theory of loss prediction [Gollakota et al., 2025] and multicalibration [Hébert-Johnson et al., 2018]. Throughout we adopt the binary-label conventions of Section 3.3. Extensions to multiclass losses likewise follow by one-vs-rest reduction.

E.1 Loss-Prediction Preliminaries

Let $\ell(h(x), y) = \mathbb{I}\{h(x) \neq y\}$ denote the 0-1 loss of a fixed classifier h. A *loss predictor* $\text{LP} : \Phi \to \mathbb{R}$ maps auxiliary features $\phi(x, h) \in \Phi$ to an estimate of $\ell(h(x), y)$. The canonical baseline is the *self-entropy predictor* $\text{SEP}(x) := \mathbb{E}[\ell(h(x), y) \mid h(x)]$ (which equals $\min\{p, 1-p\}$ for probabilistic $p = h(x)$).

Definition 6 (Advantage over the self-entropy predictor)**.** The (squared-error) *advantage* of a loss predictor LP is

$$\text{Adv}(\text{LP}) := \mathbb{E}\big[(\ell - \text{SEP})^2\big] - \mathbb{E}\big[(\ell - \text{LP})^2\big]. \tag{42}$$

A positive advantage means LP forecasts the instance-wise loss better than the model itself.

Depending on ϕ, we obtain a hierarchy of predictors: prediction-only ($\phi = h(x)$), input-aware ($\phi = (h(x), x)$), and representation-aware ($\phi = (h(x), x, r(x))$); we refer to Gollakota et al. [2025] for a detailed taxonomy.

E.2 Multicalibration Background

Multicalibration is a fine-grained notion of reliability that asks not just for global calibration, but for calibration conditional on a rich class of subpopulations or features [Hébert-Johnson et al., 2018]. At a high level, a model is multicalibrated if its predicted scores match outcomes not only on average, but also across a large collection of subsets defined by auxiliary variables or internal representations.

Definition 7 (Multicalibration Error)**.** Let C be a class of weighting functions $c\colon \Phi \to [-1,1]$, and let $h\colon \mathcal{X} \to [0,1]$ be a classifier. The *multicalibration error* of h with respect to C is defined as

$$\mathrm{MCE}(C,h) \;:=\; \max_{c\in C}\Big|\,\mathbb{E}\big[(Y-h(X))\,c(\phi(X,h))\big]\Big|. \tag{43}$$

Each function $c \in C$ defines a subpopulation or slice of the input space via its support. The quantity $\mathrm{MCE}(C,h)$ measures how well the model's predicted scores $h(x)$ match the true label Y when weighted over these slices. When C consists of indicator functions over discrete demographic subgroups, small $\mathrm{MCE}(C,h)$ implies groupwise calibration. More generally, if C includes continuous or data-dependent functions (e.g., based on internal features), low multicalibration error guarantees alignment between predicted and true outcomes across a flexible set of conditions.

In our selective classification setting, $\phi(x,h)$ may include the model's output confidence, the input x, or hidden representations from the network. The class C can be constructed accordingly to enforce calibration in feature-dependent or risk-sensitive regions of the input space.

E.3 Loss Prediction $\Longleftrightarrow$ Multicalibration

We now describe how the ability to predict one's own 0–1 loss is deeply connected to multicalibration. This perspective stems from the work of Gollakota et al. [2025], who characterize when a model "knows its own loss" in terms of multicalibration violations.

Let F be a class of loss predictors $\mathrm{LP}\colon \phi(x,h) \mapsto \hat{\ell} \in [0,1]$, which estimate the 0–1 loss $\ell(h(x),y) = \mathbb{I}\{h(x) \neq y\}$ of a fixed classifier h. As discussed in Section E.1, a loss predictor is considered good if it has a significant squared-error advantage over the model's self-estimate $\mathrm{SEP}(x)$.

Remarkably, Gollakota et al. [2025] show that this predictive advantage is tightly characterized by the multicalibration error of the model—measured over a derived weight class C that depends on the predictors in F. The following theorem formalizes this connection:

Theorem 3 (Gollakota et al. [2025], Thm. 4.1—adapted)**.** For any function class F of loss predictors and the associated weight class $C = \{(f-\mathrm{SEP})\cdot H'_\ell(h(x)) : f \in F\}$,

$$\tfrac{1}{2}\max_{\mathrm{LP}\in F}\mathrm{Adv}(\mathrm{LP}) \;\le\; \mathrm{MCE}(C,h) \;\le\; \sqrt{\max_{\mathrm{LP}\in F'}\mathrm{Adv}(\mathrm{LP})}, \tag{44}$$

where F' augments F with linear mixtures of SEP and elements of F. Thus a non-trivial advantage is possible *iff* h exhibits a multicalibration violation of similar magnitude.

This result bridges two domains: learning to predict loss (a regression task) and satisfying a generalization constraint (calibration under distributional conditions). In the selective classification setting, this insight underpins Corollary 1, which shows that the ranking error—and hence the gap to oracle performance—is tightly controlled by the model's ability to forecast its own mistakes.

E.4 Bounding the Ranking-Error Term $\varepsilon_{\mathbf{rank}}(c)$

Theorem 3 translates into a bound on the ranking error that drives the selective-classification gap.

Corollary 1 (Loss-prediction advantage controls mis-ranking)**.** Fix coverage $c \in (0,1]$ and let $\mathrm{Adv}^\star := \max_{\mathrm{LP}\in F}\mathrm{Adv}(\mathrm{LP})$ for some input-aware class F. Then the ranking-error term in Theorem 1 satisfies $\varepsilon_{\mathrm{rank}}(c) \;\le\; \sqrt{2\,\mathrm{Adv}^\star}$.

Proof. Recall that $A_c^\star = \{x : \eta_h(x) \text{ is in the top } c\text{-mass}\}$ and $A_c = \{x : g(x,h) \ge t_c\}$. Write the *difference indicator* $\delta_c(x) := \mathbb{I}_{A_c^\star}(x) \;-\; \mathbb{I}_{A_c}(x) \in \{-1,0,1\}$ so $\Pr(\delta_c = 1) = \Pr(\delta_c = -1) = c$ and $\mathbb{E}[\delta_c] = 0$.

Step 1: Express ranking error as a covariance. With $r(x) := \mathbb{I}\{h(x) = Y\}$ we have

$$\varepsilon_{\text{rank}}(c) = \mathbb{E}[r \mid A_c^\star] - \mathbb{E}[r \mid A_c] = \frac{1}{c}\,\mathbb{E}\big[r(X)\,\delta_c(X)\big]. \tag{45}$$

Step 2: Replace correctness by residual $Y - h(X)$. Because $r = 1 - \ell$ and $\ell = (Y-h)^2$ for binary labels,

$$r\,\delta_c = \big(1 - (Y-h)^2\big)\delta_c = -(Y-h)\,\delta_c \quad (\text{since } \mathbb{E}[\delta_c] = 0). \tag{46}$$

Hence

$$\varepsilon_{\text{rank}}(c) = \frac{1}{c}\,\big|\mathbb{E}[(Y - h(X))\,\delta_c(X)]\big|. \tag{47}$$

Step 3: Bound the covariance by multicalibration error. Define the bounded weight function $c^\star(x) := \delta_c(x)$; then $|c^\star(x)| \le 1$, so $c^\star \in C$ (the weight class in Theorem 3). By definition of multicalibration error,

$$\big|\mathbb{E}[(Y - h(X))\,c^\star(X)]\big| \;\le\; \text{MCE}(C, h). \tag{48}$$

Combining (47) and (48) with $c \le 1$ yields

$$\varepsilon_{\text{rank}}(c) \;\le\; \text{MCE}(C, h). \tag{49}$$

Step 4: Invoke the loss-prediction bound. Theorem 3 states $\text{MCE}(C,h) \le \sqrt{\max_{\text{LP}\in F'} \text{Adv}(\text{LP})}$. Since $F \subseteq F'$ and $\sqrt{\cdot}$ is monotone, we finally have

$$\varepsilon_{\text{rank}}(c) \;\le\; \sqrt{2\,\text{Adv}^\star}, \tag{50}$$

where the factor 2 absorbs the two-sided $F \leftrightarrow F'$ constant in Theorem 3. □

Interpretation. Let $\epsilon^2 := \max_{\text{LP}\in F} \text{Adv}(\text{LP})$ be an upper bound on loss-prediction advantage. If no loss predictor can beat self-entropy by more than ϵ^2, then the selective classifier is within $O(\epsilon)$ of the oracle at *every* coverage level. Conversely, a large loss-prediction advantage is a certificate of poor ranking and therefore of a wide gap $\Delta(c)$.

Takeaway. Loss prediction and multicalibration offer a principled lens on selective prediction: *if you cannot beat your own self-entropy predictor, you are already close to the oracle frontier*. Otherwise, the loss predictor pinpoints exactly which inputs are being mis-ranked and by how much, providing both a diagnostic and a blueprint for tightening the selective-classification gap.

E.5 Empirical Evaluation

To illustrate and validate our gap-decomposition framework, we compared four selective-classification strategies on CIFAR-10, CIFAR-100, and StanfordCars:

- `MSP`: standard maximum-softmax-probability abstention.
- `TEMP`: `MSP` with post-hoc temperature scaling.
- `SAT`: self-adaptive training, which co-trains an abstain class.
- `DE`: a deep ensemble of five `MSP` models.

For each method, we first trained a ResNet-18 on 80% of the training set (using the usual data augmentations and a held-out 20% for LP fitting). At each epoch we then:

1. Extract the 512-dim "penultimate" feature vector $\phi(x)$ from the ResNet backbone (or its ensemble average).
2. Compute the model's *self-entropy* score

$$\text{SEP}(x) \;=\; 1 - \max_j\; p_j(x) \quad \text{with} \quad p_j(x) = \text{softmax}_j(\text{logits}(x)/T).$$

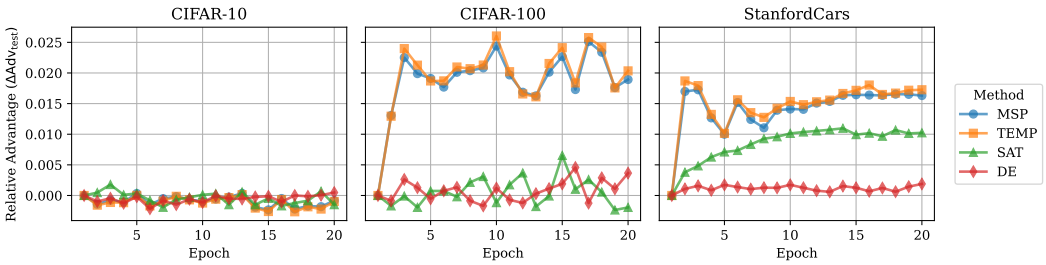


Figure 6: **Relative LP advantage over training epochs across datasets**. For each method, we plot the shift in test-set advantage $\Delta\mathrm{Adv}_{\mathrm{test}}(t)$ relative to epoch 1, indicating how much additional ranking signal the loss predictor learns over time. Larger values imply greater misalignment between the model's confidence and correctness.

3. Train a small MLP LP: $\phi(x) \mapsto \widehat{\ell} \in [0,1]$ to minimize $\mathbb{E}\big[\big(\widehat{\ell} - \mathbb{I}\{\hat{y}(x) \neq y\}\big)^2\big]$ on the held-out 20% split.
4. Measure the *LP advantage* on the *test* set,

$$\mathrm{Adv}_{\mathrm{test}} = \mathbb{E}\big[(\ell - \mathrm{SEP})^2\big] - \mathbb{E}\big[(\ell - \mathrm{LP})^2\big], \quad \ell = \mathbb{I}\{\hat{y}(x) \neq y\},$$

and record its shift relative to the first epoch $\Delta\mathrm{Adv}_{\mathrm{test}}(t) = \mathrm{Adv}_{\mathrm{test}}(t) - \mathrm{Adv}_{\mathrm{test}}(1)$.

Loss–Prediction Network. Below is the PyTorch representation of our two-hidden-layer LP head. It takes the ResNet features (optionally concatenated with SEP) and regresses the per-example 0–1 loss via mean-squared error.

```
LossPredictor(
  (net): Sequential(
    (0): Linear(in_features=512, out_features=128, bias=True)
    (1): ReLU()
    (2): Dropout(p=0.5)
    (3): Linear(in_features=128, out_features=64, bias=True)
    (4): ReLU()
    (5): Dropout(p=0.5)
    (6): Linear(in_features=64, out_features=1, bias=True)
  )
)
```

Key observations. On CIFAR-10 (left panel of Figure 6), all methods stay close to zero $\Delta\mathrm{Adv}_{\mathrm{test}}$, indicating that the model's own confidence scores already capture most of the available ranking signal. On CIFAR-100 (middle panel), `MSP` and `TEMP` exhibit large positive shifts in LP advantage, suggesting that a dedicated loss predictor can substantially improve ranking—consistent with a larger gap from the oracle. By contrast, `SAT` and `DE` remain near zero, indicating that their confidence scores are already well aligned with correctness. On StanfordCars (right panel), the gap widens even further: both `MSP` and `TEMP` allow for significant gains via loss prediction, and even `SAT` leaves nontrivial room for improvement. Only `DE` consistently resists such gains, implying that deep ensembling is uniquely effective at preserving reliable ranking in high-variance domains.

Conclusion. These results match our theory perfectly: whenever the LP head cannot improve on self-entropy, the selective classifier is effectively oracle-optimal; whenever it can, the size of that advantage precisely quantifies the remaining ranking error and the gap from the ideal frontier.

F Additional Results

F.1 Calibration Experiments

E-AURC vs ECE We provide additional comparisons on more datasets (CIFAR-10 and StanfordCars) on the relationship between the selective classification gap and the model's expected calibration

error. See Tables 2 and 3 for exact results. In general, our conclusions from Section 4.2 hold here as well: while temperature scaling (TEMP) improves ECE over MSP, it does not reduce the selective classification gap—underscoring the limits of monotone calibration. In contrast, SAT and deep ensembles (DE) improve both ECE and gap by altering the ranking, confirming that only re-ranking methods yield meaningful gains in selective performance.

Table 2: **Experiments on calibration across model classes on CIFAR-10**. Similar as Table 1

	CNN				ResNet-18				WideResNet-50			
	MSP	TEMP	SAT	DE	MSP	TEMP	SAT	DE	MSP	TEMP	SAT	DE
E-AURC	0.024	0.023	0.019	0.016	0.004	0.004	0.003	0.002	0.003	0.003	0.002	0.002
ECE	0.075	0.025	0.035	0.010	0.025	0.014	0.016	0.007	0.027	0.022	0.019	0.010

Table 3: **Experiments on calibration across model classes on StanfordCars**. Similar as Table 1

	CNN				ResNet-18				WideResNet-50			
	MSP	TEMP	SAT	DE	MSP	TEMP	SAT	DE	MSP	TEMP	SAT	DE
E-AURC	0.176	0.177	0.166	0.159	0.030	0.029	0.26	0.022	0.026	0.026	0.23	0.020
ECE	0.110	0.025	0.058	0.025	0.040	0.027	0.037	0.025	0.017	0.017	0.015	0.015

F.2 Extension to Large Language Models

Our primary experiments focus on vision and synthetic datasets, where uncertainty and selective prediction have well-established definitions and evaluation metrics. In these domains, notions such as confidence calibration, abstention rates, and oracle coverage curves provide a clear framework for measuring reliability. Extending the same analysis to large language models, however, presents new difficulties as outlined in particlar by the following two challenges:

1. **Uncertainty for generative models remains ill-defined.** Even for classification-style prompts, the community has not fully converged on how to translate sequence-level probabilities into abstention scores.
2. **Prompting artefacts add variance.** Small changes in in-context examples or decoding settings can swamp the effects we wish to isolate.

Despite these challenges, we have added a focused set of LLM experiments to demonstrate that our five-term decomposition still diagnoses the gap.

F.2.1 Approximation Error — Scaling from 4B $\rightarrow$ 12B

We evaluate Gemma 3-IT 4B and 12B [Team et al., 2025] on ARC-Challenge (ARC-C, 25-shot) [Clark et al., 2018] and MMLU (5-shot, top-1) [Hendrycks et al., 2021] using the standard MSP score on the *first answer token* (no further fine-tuning).

Table 4: Accuracy comparison across model scales.

Model	**ARC-C Accuracy**	**MMLU Accuracy**
Gemma 3-IT 4B	56.2%	59.6%
Gemma 3-IT 12B	68.9%	74.5%

Observation. Consistent with our vision experiments, increasing capacity reduces the gap.

F.2.2 Bayes Error — Separating Easy vs. Noisy Questions

Following the MMLU-Pro protocol [Wang et al., 2024], we partition the validation set into the *easiest 25%* and *noisiest 25%* questions (based on human–LLM agreement).

Table 5: Selective-classification gap area (lower is better).

Model	ARC-C E-AURC	MMLU E-AURC
Gemma 3-IT 4B	0.114	0.107
Gemma 3-IT 12B	0.091	0.082

Table 6: E-AURC across data difficulty levels on Gemma 3-IT 4B.

Split	E-AURC
Full MMLU	0.107
Easiest quartile	0.018
Noisiest quartile	0.316

Observation. When intrinsic Bayes noise is low (easy questions), the gap nearly vanishes; when noise is high, the gap widens.

F.2.3 Ranking Quality — Calibration vs. Re-ranking

We keep the backbone fixed on Gemma 3-IT 4B and compare the ranking quality of the following uncertainty scores:

- `MSP`,
- `TEMP` (scalar T fitted on a held-out validation split),
- `DE` of five LoRA-fine-tuned replicas.

Table 7: Calibration and gap performance across ranking methods on Gemma 3-IT 4B.

	ARC-C			MMLU		
	`MSP`	`TEMP`	`DE`	`MSP`	`TEMP`	`DE`
E-AURC	0.127	0.126	0.087	0.122	0.122	0.079
ECE	0.171	0.092	0.056	0.135	0.084	0.052

Observation. Temperature scaling lowers ECE yet leaves the gap untouched; the ensemble both calibrates *and* improves ranking, shrinking the gap.

Summary. These results confirm that our decomposition extends to LLMs: capacity, Bayes noise, and ranking quality each contribute measurable terms. A full generative-text study (e.g., free-form question answering or code synthesis) will require new abstention semantics and we leave a more thorough treatment for future work.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Our abstract and intro reflects the contributions accurately.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: See Section 5.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: We provide a short proof in the main paper and a more extensive analysis and proof in Appendix B.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: See Appendix D.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: We include our full experimental suite and details for reproducibility.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: See Appendix D.2.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: All of our reported results are reported as mean values over 5 random runs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: See Appendix D.1.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The paper conforms to the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: See Appendix A.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: We are not releasing any new assets that require any specific safeguards.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We have cited related work appropriately.

    Guidelines:

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

13. **New assets**

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

Answer: [Yes]

Justification: We are releasing our codebase to aid reproducibility.

Guidelines:

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

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

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

Answer: [NA]

Justification: We did not conduct any crowdsourcing and/or research with human subjects.

Guidelines:

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

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

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

Answer: [NA]

Justification: We did not conduct any user studies requiring IRB approval.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [NA]

Justification: We only used LLMs to help with paper editing.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.