Ensemble Selective Classification

Sinian Zhang[1*] **Chongwei Chen**[2*] **Guanchen Li**[2,3] **Ju Sun**[2]
[1]Division of Biostatistics and Health Data Science, University of Minnesota, Minneapolis, MN
[2]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN
[3]Department of Mathematics and Computer Science, Bethel University, St. Paul, MN
{zhan9381,chen8596,li004004,jusun}@umn.edu

Project page: https://github.com/sun-umn/Ensemble_SC

Abstract

Selective classification (SC) improves model reliability by allowing a classifier to abstain on potentially wrong predictions. In the practical post-hoc setting, where the base classifier is fixed and often accessed as a black box, many confidence scores have been proposed, yet no single score is consistently best across models, datasets, and deployment shifts. We address this limitation by proposing, to our knowledge, the first ensemble framework for post-hoc SC, which learns to combine a diverse set of existing confidence scores using a small labeled calibration set drawn from the deployment distribution. We cast this problem as learning to separate correct from incorrect predictions of the base classifier, and train the ensemble selector with SoftRank-AURC, a differentiable soft-rank plug-in objective for the area under the risk-coverage curve (AURC), the standard evaluation metric for SC. We also study the SElective classifier LEarning (SELE) loss as an alternative training objective. For SoftRank-AURC, our analysis suggests that the learned ensemble may be competitive with a strong individual scorer, up to estimation error and smoothing bias. Across 15 image and NLP classification tasks, with additional LLM multiple-choice experiments, the proposed method is consistently strong and often outperforms the best individual score, particularly under distribution shift, establishing ensemble post-hoc SC as a simple and effective way to improve reliability without retraining the base classifier.

1 Introduction

Selective classification (SC) allows a classifier to abstain from acting on samples it potentially misclassifies, thereby ensuring high accuracy on the examples it actually classifies [1–3]. The core of SC relies on a confidence scorer that estimates the likelihood of a prediction being correct. By setting a rejection threshold on the confidence score, the model decides whether to confidently output a label or strategically abstain, directly trading off coverage (the portion of predictions made) for reliability (the accuracy of those predictions).

Current SC methods can be divided into **in-training** and **post-hoc** approaches: the former jointly optimizes a confidence scorer with the classifier, and the latter works directly with pretrained classifiers [4, 3, 5]. **In this paper, we focus on the post-hoc approach** because (1) it is essential for classifiers deployed as black-box services, where the training process is inaccessible; and (2) the deployment setting can often be different from the training setting, and the resulting distribution shift can substantially degrade in-training SC methods [4, 3, 6].

*Equal contribution.

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

Most post-hoc SC scorers are hand-crafted. However, it is well understood that the performance of any confidence scorer depends on the combination of (classifier, data distribution) with which it is involved, and hence there is **no universally optimal hand-crafted confidence scorer** [4, 3]. This makes single-scorer SC deployment brittle, because real-world distribution shifts are common and often change over time [7, 8].

In this paper, instead of designing another scorer, **we propose an ensemble SC framework that synergizes multiple confidence scorers to improve SC performance**. Although ensemble learning is classic in machine learning [9, 10] and the idea of learning to combine multiple scorers seems natural, surprisingly, to the best of our knowledge, **ours is the first in the context of post-hoc SC**. To facilitate ensemble learning, we assume access to a reasonably sized calibration set whose data points follow the deployment (i.e., test) distribution—hence **our ensemble SC framework allows distribution shifts between the deployment and the training settings**. With the base classifier fixed, labeling the calibration set based on whether each prediction is correct or not leads to a binary correctness-prediction problem. Our technical contributions include

- **Proposing** an ensemble SC framework to boost post-hoc SC performance by learning to combine the strengths of multiple base post-hoc confidence scorers. Based on a calibration set, the learning problem is framed as post-hoc correctness prediction (Sec. 3.1);
- **Developing** *SoftRank-AURC*, our primary training objective for learning an optimal ensemble of the base scorers. Based on pairwise soft-rank, it is a convenient finite-sample approximation to the AURC metric friendly for gradient-based optimization (Sec. 3.2);
- **Deriving** generalization guarantees for the SoftRank-AURC objective, showing that the learned ensemble performs competitively with the best base scorer, up to estimation and smoothing errors under regularity assumptions (Sec. 3.3);
- **Conducting** systematic evaluation of our proposed ensemble SC methods against 10 popular base scorers, on 15 image and NLP classification tasks as well as LLM-based multiple-choice problems, demonstrating the superiority of our ensemble SC (Sec. 4).

2 Technical Background & Related Work

Basic SC setup and evaluation metrics For any natural number N, we define $[N] \doteq \{1, \dots, N\}$. Consider a K-class classification problem with its input space $\mathcal{X} \subseteq \mathbb{R}^n$ and label space $\mathcal{Y} = [K]$. A selective classifier (f, g) consists of a classifier $f : \mathcal{X} \to \mathbb{R}^K$ and a selector $g : \mathcal{X} \to \{+1, -1\}$. We write $f^\ell(\boldsymbol{x}) \doteq \operatorname{argmax}_{k \in [K]} f_k(\boldsymbol{x})$ for the class label predicted by f under the standard argmax rule. The selective classifier works as follows:

$$(f, g)(\boldsymbol{x}) \doteq \begin{cases} f^\ell(\boldsymbol{x}) & \text{if } g(\boldsymbol{x}) = +1 \\ \text{abstain} & \text{if } g(\boldsymbol{x}) = -1 \end{cases}, \tag{2.1}$$

for any input $\boldsymbol{x} \in \mathcal{X}$. Typical selectors g take a confidence-thresholding form

$$g_{s,\gamma}(\boldsymbol{x}) \doteq 2a_{s,\gamma}(\boldsymbol{x}) - 1, \quad \text{with } a_{s,\gamma}(\boldsymbol{x}) \doteq \mathbb{I}(s(\boldsymbol{x}) \geq \gamma), \tag{2.2}$$

where $s(\cdot)$ is a confidence *scorer*, and γ is a tunable selection threshold. For a given selective classifier $(f, g_{s,\gamma})$, its SC performance on a deployment distribution $\mathcal{D}^{\text{depl}}_{\mathcal{X},\mathcal{Y}'}$ (or its abbreviated form depl) is often measured jointly by the *coverage* and the selective *risk* defined as follows:

$$\phi_{s,\gamma} = \mathbb{E}_{\text{depl}}\left[a_{s,\gamma}(\boldsymbol{x})\right] \text{ (\textbf{coverage})}, \quad R_{s,\gamma} = \mathbb{E}_{\text{depl}}\left[\ell(f^\ell(\boldsymbol{x}), y) a_{s,\gamma}(\boldsymbol{x})\right] / \phi_{s,\gamma} \text{ (\textbf{risk})}. \tag{2.3}$$

In other words, coverage is the proportion of predictions selected (i.e., accepted), and (selective) risk is the error rate on the selected subset. We note that $\mathcal{D}^{\text{depl}}_{\mathcal{X},\mathcal{Y}'}$ **does not have to be the same as the training distribution, to allow distribution shifts**, following the generalized SC setup in [4]; in particular, $\mathcal{Y}'$ can be different from $\mathcal{Y}$ to allow label shifts that are commonly considered in the out-of-distribution (OOD) detection literature. Most of the existing work sets the ℓ to be the standard $0/1$ error [11–13], and we follow it in this paper. SC is usually evaluated through the whole spectrum of risk-coverage tradeoff, and a popular summarizing metric is the area-under-the-risk-coverage-curve (AURC), $\text{AURC}(f, s) = \int_0^1 R_{s,\gamma}(\phi_{s,\gamma})\, d\phi_{s,\gamma}$. Since a lower risk is favored for any fixed coverage, a lower AURC indicates a better overall tradeoff.

Post-hoc SC vs. in-training SC In-training SC performs joint classifier-selector training, whereas post-hoc SC works with given classifiers and does not access the training process. As argued in

Sec. 1, **in this paper we focus on post-hoc SC** because it is more practical in black-box deployment settings; we refer the reader to [4, 3, 5] for a review of in-training SC. In post-hoc SC, the classifier f typically provides the *raw logits* $\boldsymbol{z} \doteq f(\boldsymbol{x}) \in \mathbb{R}^K$ for any $\boldsymbol{x} \in \mathcal{X}$; intermediate features, if available, can be used as additional confidence signals. A typical post-hoc SC pipeline as summarized in Alg. 1; We assume that $\mathcal{D}^{\text{cali}}_{\mathcal{X},\mathcal{Y}'}$ is the same as $\mathcal{D}^{\text{depl}}_{\mathcal{X},\mathcal{Y}'}$.

Algorithm 1 A typical post-hoc SC pipeline [4]

Input: A given classifier f; a scorer s; a calibration dataset $Z^{\text{cali}} \sim_{iid} \mathcal{D}^{\text{cali}}_{\mathcal{X},\mathcal{Y}'}$; a risk/coverage target

1: $\forall(\boldsymbol{x}_i, y_i) \in Z^{\text{cali}}$, compute $s(\boldsymbol{x}_i)$ and $\mathbb{I}(f^\ell(\boldsymbol{x}_i) \neq y_i)$, and sort them by confidence
2: Determine γ on Z^{cali} according to the risk/coverage target
3: Deploy the selector $g_{s,\gamma}$ based on Eq. (2.2).

Many classic hand-crafted scorers are defined on softmax responses (SRs) $\text{softmax}(\boldsymbol{z})$. The most popular is the maximum softmax probability $\text{SR}_{\max}(\boldsymbol{z}) \doteq \max_{k\in[K]}[\text{softmax}(\boldsymbol{z})]_k$ [14, 1]. Although simple, $\text{SR}_{\max}$ is surprisingly effective and often outperforms more complex in-training ones [15]. However, these SRs-based scorers generally assume that SRs approximate posterior probabilities well [2], which often does not hold in practice. To address this, prior work proposes margin-based scorers such as the confidence margin ($\text{RL}_{\text{conf-M}}$) and the geometric margin ($\text{RL}_{\text{geo-M}}$) that operate directly on raw logits $\boldsymbol{z}$ [4]. Beyond these, other widely used post-hoc scorers include the entropy of the SRs [14], energy [16], feature-space proximity (e.g., kNN distance on normalized penultimate features) [17], representation rectification (ReAct) [18], and ViM/ViM-residual from subspace projections [19]. Clearly, the performance of any given scorer depends on $(f, \mathcal{D}^{\text{depl}}_{\mathcal{X},\mathcal{Y}'})$, and recent systematic evaluations, e.g., [4, 3], strongly suggest that **there is no universal winner**.

AURC-related computation and learning Recent work has studied AURC from several complementary angles. [2] proposes the SElective classifier LEarning (SELE) loss as a smooth approximation to AURC and performs post-hoc scorer learning. [20] derives finite-sample plug-in estimators for the population-level AURC based on empirical ranks, and finetunes pretrained classifiers to improve SC performance based on $\text{SR}_{\max}(\boldsymbol{z})$. Similar to them, here we also take AURC as a learning objective to learn effective scorers, thereby improving the final SC performance; different from them, here our learning is based on a learnable ensemble of multiple given scorers/selectors, rather than learning from scratch or finetuning the classifier with a fixed scorer. Further away, [21] connects confidence calibration and SC in low-confidence regions and proposes a modified AURC loss to improve calibration; [5] boosts the selective prediction performance of large language models (LLMs) by finetuning them with respect to the AURC loss.

Ensemble methods in general & for SC Ensemble methods, i.e., combining multiple machine learning models to improve performance, are a cornerstone of modern machine learning [9, 10]. In the context of uncertainty quantification and SC, **deep ensembles** methods have emerged as a powerful and widely used technique for producing confidence or uncertainty scorers [22, 23]. However, because they require multiple trained models or variants of the same trained model, they are not directly applicable to the *single-fixed-classifier, black-box post-hoc setting* studied here. Moreover, prior work combines in-distribution SC scorers with out-of-distribution (OOD) detection scorers for simultaneous SC and OOD detection [24, 25], but optimizes the SC and OOD detection objectives separately. Here, we follow the generalized SC setup proposed in [4]—which covers both classic SC and OOD detection, and combine various base scorers to optimize the SC performance.

3 Our Method: Ensemble Selective Classification

Since there is no universal winner among post-hoc confidence scorers for SC, our goal here is to derive an ensemble SC framework that combines the strengths of multiple post-hoc confidence scorers and works uniformly better than any individual scorer across different classifiers and deployment distributions. We first set up the problem in Sec. 3.1, then introduce our ensemble scorers and training algorithms in Sec. 3.2, and finally provide theoretical analysis in Sec. 3.3.

3.1 Problem formulation

To set up the problem, we assume the basic generalized SC setting in Sec. 2. Moreover, we assume T base post-hoc scorers $\tilde{s}_1, \ldots, \tilde{s}_T$, each $\mathcal{X} \to \mathbb{R}$. **Our task is to learn an ensemble of the T base scorers**. To facilitate learning, we need a calibration set: let $Z^{\text{cali}} \doteq \{(\boldsymbol{x}_i, y_i)\}_{i\in[N]}$ be a labeled calibration set iid sampled from the deployment distribution, i.e., $\mathcal{D}^{\text{depl}}_{\mathcal{X}\times\mathcal{Y}'} = \mathcal{D}^{\text{cali}}_{\mathcal{X}\times\mathcal{Y}'}$ for our case. The classifier's predictions $f^\ell(\boldsymbol{x}_i)$'s induce binary correctness labels

$$y_i^z \doteq \begin{cases} +1 & f^\ell(\boldsymbol{x}_i) = y_i \quad \text{(correctly classified)} \\ -1 & f^\ell(\boldsymbol{x}_i) \neq y_i \quad \text{(wrongly classified)} \end{cases}, \tag{3.1}$$

leading to a binary-labeled calibration set $Z_B^{\text{cali}} \doteq \{(\boldsymbol{x}_i, y_i^z)\}_{i\in[N]}$. To ensure that our ensemble scorer has nontrivial gains over the best individual scorer, we assume without formal quantification that the base scorers are reasonably diverse, and a nontrivial subset of them are reasonably accurate.

Scorer normalization Before combining the base scorers, we normalize each scorer by the following robust procedure: for each $t \in [T]$, let

$$m_t \doteq \text{median}\{\tilde{s}_t(\boldsymbol{x}_i) : i \in [N]\}, \qquad \text{MAD}_t \doteq \text{median}\{|\tilde{s}_t(\boldsymbol{x}_i) - m_t| : i \in [N]\}, \tag{3.2}$$

where MAD_t denotes the median absolute deviation (MAD). The normalized scorer is defined as

$$s_t(\boldsymbol{x}) \doteq \sigma[(\tilde{s}_t(\boldsymbol{x}) - m_t)/b_t] \in (0,1) \quad \text{with } b_t \doteq \max\{c_{\text{MAD}}\text{MAD}_t, \delta_{\min}\}, \tag{3.3}$$

where $\sigma(\cdot)$ is the sigmoid function, and $c_{\text{MAD}}, \delta_{\min} > 0$ are tunable hyperparameters. The median and MAD provide robust estimates of each scorer's center and scale; we provide the complete motivation and additional details for the procedure in App. A.1. Note that the transformation defined in Eq. (3.3) is monotonically increasing in $\tilde{s}_t(\boldsymbol{x})$, so it preserves the confidence ranking across samples per scorer while transforming all scorers' output into the common range $(0, 1)$.

3.2 Ensemble scorers and our algorithm

Ensemble scorers The simplest choice is a direct linear combination of the normalized base scorers, denoted as S_{scr}:

$$S_{\text{scr}}(\boldsymbol{x}; \boldsymbol{\omega}) \doteq \sum\nolimits_{t\in[T]} \omega_t s_t(\boldsymbol{x}). \tag{3.4}$$

However, even after the robust normalization, the empirical scores produced by the different scorers may still have different distributions. Linear combinations of numerical values from these heterogeneous distributions tend to cause biases and undermine the overall performance. Thus, our main ensemble scorer is a **linear combination of learnable selectors**, denoted as S_{slr}:

$$\boxed{S_{\text{slr}}(\boldsymbol{x}; \boldsymbol{\omega}, \boldsymbol{\gamma}) \doteq \sum\nolimits_{t\in[T]} \omega_t \,\text{sign}(s_t(\boldsymbol{x}) - \gamma_t)} \tag{3.5}$$

with $\boldsymbol{\omega} \geq 0$ and $\boldsymbol{\gamma} \in (0,1)^T$. Here, inside $S_{\text{slr}}(\boldsymbol{x}; \boldsymbol{\omega}, \boldsymbol{\gamma})$, the learnable scorer-adaptive thresholds γ_t's can cancel out potential scorer-level mode differences in distribution, and the $\text{sign}(\cdot)$ operation puts all scores on the same numerical footing. So, $S_{\text{slr}}(\boldsymbol{x}; \boldsymbol{\omega}, \boldsymbol{\gamma})$ is expected to have more robust and likely superior performance than $S_{\text{scr}}(\boldsymbol{x}; \boldsymbol{\omega})$—confirmed in our experiments (Sec. 4). Moreover, $S_{\text{slr}}(\boldsymbol{x}; \boldsymbol{\omega}, \boldsymbol{\gamma})$ can also be viewed as a boosting-style model whose base selectors $\text{sign}(s_t(\boldsymbol{x}) - \gamma_t)$'s are decision stumps [26].

Why not more complex combinations? One might ask why we do not consider more complex combinations, in particular, why we do not learn a scorer directly from the raw or softmax logits, as most of the base post-hoc scorers we consider are functions of the raw/softmax logits. **Our main consideration here is data efficiency**: setting up a simple multi-layer perceptron (MLP) to learn a scorer directly on the raw/softmax logits requires $O(K^2)$ learnable parameters when its hidden-layer widths are proportional to K, the number of classes; while both $S_{\text{slr}}(\boldsymbol{x}; \boldsymbol{\omega}, \boldsymbol{\gamma})$ and $S_{\text{scr}}(\boldsymbol{x}; \boldsymbol{\omega})$ only have $O(T)$ learnable parameters: T weights for SCR, and T weights plus T thresholds for SLR. Since in practice $T \ll K$, we prefer the lightweight solution with $S_{\text{slr}}(\boldsymbol{x}; \boldsymbol{\omega}, \boldsymbol{\gamma})$ and $S_{\text{scr}}(\boldsymbol{x}; \boldsymbol{\omega})$ that likely only need small amounts of calibration data—especially useful when data are scarce and/or the deployment distribution can drift over time and thus frequent re-calibration might be necessary.

Our training loss and algorithm With the binary-labeled validation set Z_B^{cali} and the learnable ensemble scorers S_{slr} and S_{scr}, we face a standard binary classification problem. It is tempting to use the classic surrogate losses for binary classification, such as cross-entropy or hinge loss, to train the ensemble scorers. However, these losses are not directly related to AURC, the metric of primary interest for SC. Thus, **we set out to perform training by optimizing the AURC metric directly**.

We start with the following characterization of the population AURC, which takes the form of a rank-weighted risk

$$\text{AURC}(f, s) = \mathbb{E}_{\boldsymbol{x}, y^z}[-\log(1 - \mathbb{E}_{\boldsymbol{x}'}\mathbb{I}\{s(\boldsymbol{x}') \leq s(\boldsymbol{x})\})\,\mathbb{I}\{y^z = -1\}]\,. \tag{3.6}$$

This characterization first appears in [20]; we provide a self-contained and much simplified derivation in App. A.2. On finite datasets, both the outer and the inner expectations can be approximated by the empirical averages over Z_B^{cali}, yielding

$$\text{AURC}(f, s) \approx \frac{1}{N}\sum\nolimits_{i=1}^{N} \alpha_i(s)\mathbb{I}\{y_i^z = -1\}, \quad \text{where } \alpha_i(s) \doteq -\log\left(1 - \frac{1}{N}\sum\nolimits_{j=1}^{N}\mathbb{I}\{s(\boldsymbol{x}_j) \leq s(\boldsymbol{x}_i)\}\right). \tag{3.7}$$

Moreover, both α_i and our ensemble scorer S_{slr} contain indicator functions, which have zero gradient almost everywhere. Hence, to optimize the empirical AURC loss with gradient-based methods, we smooth the two types of indicator functions as follows:

$$\mathbb{I}\{s(\boldsymbol{x}_j) \leq s(\boldsymbol{x}_i)\} \longrightarrow \sigma\left(\frac{s(\boldsymbol{x}_i) - s(\boldsymbol{x}_j)}{\kappa}\right) \quad \text{with } \sigma(u) \doteq (1 + e^{-u})^{-1} \quad \forall i, j, \tag{3.8}$$

$$S_{\text{slr}} \longrightarrow S_{\text{slr}}^{\beta}(\boldsymbol{x}; \boldsymbol{\omega}, \boldsymbol{\gamma}) \doteq \sum\nolimits_{t\in[T]} \omega_t \tanh[(s_t(\boldsymbol{x}) - \gamma_t)/\beta], \tag{3.9}$$

where $\kappa > 0$ and $\beta > 0$ are smoothing parameters for the sigmoid and tanh functions, respectively. Then, we arrive at our final training loss:

$$\boxed{\mathcal{L}_A(f, S_{\text{slr}}^{\beta}; \kappa) \doteq \frac{1}{N}\sum\nolimits_{i=1}^{N} \widehat{\alpha}_i(S_{\text{slr}}^{\beta}; \kappa)\,\mathbb{I}\{y_i^z = -1\}}, \quad \textbf{(SoftRank-AURC)} \tag{3.10}$$

where $\widehat{\alpha}_i(S_{\text{slr}}^{\beta}; \kappa) \doteq -\log\left[1 - \frac{1}{N}\sum_{j=1}^{N}\sigma\left(\frac{1}{\kappa}(S_{\text{slr}}^{\beta}(\boldsymbol{x}_i) - S_{\text{slr}}^{\beta}(\boldsymbol{x}_j))\right)\right]$. We train our plain weighted scorer model S_{scr} similarly, without the inner smoothing.

Comparison with alternative losses. The SELEctive classifier LEarning (SELE) loss has been proposed as a computationally manageable proxy of AURC, with its indicator version tightly approximating AURC and its smooth version being Fisher consistent for learning a proper uncertainty score [2]. That convention ranks larger scores as more uncertain, whereas our ensemble score is a confidence score. We therefore reverse the score direction in the pairwise surrogate: an erroneous point i should not receive higher confidence than another point j, leading to

$$\mathcal{L}_S \doteq \tfrac{1}{N^2}\sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{N}\mathbb{I}\{y_i^z = -1\}\,\log(1 + \exp(S(\boldsymbol{x}_i) - S(\boldsymbol{x}_j)))\,. \tag{3.11}$$

The term is large when an incorrectly classified point i receives higher confidence than many other calibration points, exactly the failure mode that increases the early-coverage portion of the RC curve. We optimize $\mathcal{L}_S$ over the same ensemble score family using the full calibration batch.

3.3 Theoretical analysis

In this section, we analyze the smooth ensemble scorer S_{slr}^{β} in Eq. (3.9) with the SoftRank-AURC objective, and provide generalization guarantees for the learned ensemble scorer. We focus on the hypothesis class

$$\mathcal{S} \doteq \left\{S_{\text{slr}}^{\beta}(\cdot; \boldsymbol{\omega}, \boldsymbol{\gamma}) \,:\, \boldsymbol{\omega} \in \Delta^T,\, \boldsymbol{\gamma} \in [0, 1]^T\right\}, \tag{3.12}$$

where $\Delta^T \doteq \{\boldsymbol{\omega} \in \mathbb{R}_+^T : \sum_{t=1}^{T}\omega_t \leq 1\}$. In other words, we control the capacity of the hypothesis class by restricting $\|\boldsymbol{w}\|_1 \leq 1$, following standard practice in statistical learning theory. One can of course choose other constants for the upper bound and follow our analysis to obtain similar results.

Algorithm 2 SoftRank-AURC Ensemble Selective Classification

Input: Oriented raw post-hoc scorers $\{\tilde{s}_t : \mathcal{X} \to \mathbb{R}\}_{t=1}^{T}$; calibration set Z_B^{cali} with labels $\{y_i^z\}$; smoothing temperatures $\beta > 0$, $\kappa > 0$; optimizer; outer iterations E.

Output: Learned weights $\boldsymbol{\omega}^*$ and inner thresholds $\boldsymbol{\gamma}^*$.

1: **Normalize scores:** Apply the score-normalization procedure in App. A.1 to compute normalized score values $\{s_t(\boldsymbol{x}_i) : t \in [T], i \in [N]\}$ on the calibration set.
2: **Initialize:** Unconstrained parameter vectors $\boldsymbol{\theta}_{\boldsymbol{\omega}}$ and $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$.
3: **for** epoch $= 1$ to E **do**
4: Compute ensemble weights: $\boldsymbol{\omega} \leftarrow \text{abs}(\boldsymbol{\theta}_{\boldsymbol{\omega}})$.
5: Compute thresholds: $\boldsymbol{\gamma} \leftarrow \sigma(\boldsymbol{\theta}_{\boldsymbol{\gamma}})$ elementwise.
6: Compute ensemble scores $S_{\text{slr}}^{\beta}(\boldsymbol{x}_i; \boldsymbol{\omega}, \boldsymbol{\gamma})$ via Eq. (3.9).
7: Compute SoftRank-AURC loss $\mathcal{L}_A$ via Eq. (3.10).
8: Update $\theta = (\boldsymbol{\theta}_{\boldsymbol{\omega}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}})$ using the optimizer on $\mathcal{L}_A$.
9: **end for**
10: **Return:** $\boldsymbol{\omega}^* = \text{abs}(\boldsymbol{\theta}_{\boldsymbol{\omega}})$ and $\boldsymbol{\gamma}^* = \sigma(\boldsymbol{\theta}_{\boldsymbol{\gamma}})$, with σ applied elementwise.

Assumption 1 (Regularity of the ensemble scorer)**.** *On the calibration distribution $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}'}^{\text{cali}}$, every ensemble scorer $S \in \mathcal{S}$ satisfies*

$$\Pr\big(a < S(\boldsymbol{x}) \le b\big) \le L_{a,b,S}\,(b-a) \text{ for some } L_{a,b,S} > 0 \quad \forall\, [a,b] \subset \mathbb{R}. \tag{3.13}$$

In other words, the probability mass of any ensemble scorer $S \in \mathcal{S}$ acting on the calibration distribution is sufficiently widespread. The supporting lemmas are deferred to App. B.2.

Theorem 2 (Generalization bounds for SoftRank-AURC-based learning)**.** *Under Assp. 1, the following holds uniformly for all constants $\beta > 0$, $\kappa \in (0, 1]$ and classifier f: $\forall \delta \in (0, 1)$, consider*

$$(\widehat{\boldsymbol{\omega}}, \widehat{\boldsymbol{\gamma}}) \in \operatorname{argmin}_{(\boldsymbol{\omega},\boldsymbol{\gamma})\in\Delta^T\times[0,1]^T} \mathcal{L}_A(f, S_{\text{slr}}^{\beta}; \kappa), \qquad \widehat{S} \doteq S_{\text{slr}}^{\beta}(\cdot; \widehat{\boldsymbol{\omega}}, \widehat{\boldsymbol{\gamma}}), \tag{3.14}$$

where the SoftRank-AURC loss $L_A(f, S_{\text{slr}}^{\beta}; \kappa)$ is as defined in Eq. (3.10). Then, with probability at least $1 - \delta$,

$$\text{AURC}(f, \widehat{S}) \le \inf_{S\in\mathcal{S}} \text{AURC}(f, S) + \widetilde{O}_{\text{hp}}\left(\sqrt{\frac{T + \log(1/\delta)}{N}}\right) + \widetilde{O}\Big(\kappa \log \frac{e}{\kappa}\Big). \tag{3.15}$$

Here, $\widetilde{O}_{\text{hp}}$ suppresses logarithmic factors in $N, 1/\beta, 1/\kappa$, and $\widetilde{O}$ hides dependency on the regularity parameters in Assp. 1. This theorem has several interesting implications: (1) The same bound in Eq. (3.15) holds with the infimum term replaced by $\min_{t\in[T],\,\gamma\in[0,1]} \text{AURC}\big(f, \boldsymbol{x} \mapsto \tanh((s_t(\boldsymbol{x}) - \gamma)/\beta)\big)$. This implies that the learned ensemble may be competitive with a strong individual scorer in terms of AURC, up to estimation error and smoothing bias. (2) The data-dependent rate $\sqrt{T/N}$ confirms that we only need N to dominate T (perhaps up to logarithmic factors) to enjoy effective learning. So, the proposed SoftRank-AURC-based framework is highly data-efficient, as in practice T is very small.

4 Experiments

We evaluate the proposed method on image/NLP datasets and LLM multiple-choice benchmarks. We also conduct ablation studies on the number of base selectors and the calibration size. Lower AURC is better throughout. The experimental comparisons are organized as follows. **(1) Fixed-classifier ensemble setting.** For each fixed base classifier, our model learns a post-hoc ensemble confidence score from normalized confidence signals. We report both the score-combiner S_{scr}, abbreviated SCR, and the selector-combiner S_{slr}^{β}, abbreviated SLR; SLR is our main structured selector because it combines thresholded base selectors rather than raw score values, reducing reliance on residual scale and distribution differences across confidence scores. **(2) Training objectives.** Our main training objective is SoftRank-AURC, which directly optimizes a smooth plug-in approximation to AURC, yielding AURC-SCR and AURC-SLR in the main tables. We also train the same two model families with SELE, an AURC-oriented pairwise surrogate, yielding SELE-SCR and SELE-SLR; the SELE

objective is defined in Eq. (3.11). **(3) Baselines and learned black-box comparison.** Following the benchmark protocol in [4], we compare against the ten post-hoc confidence scores summarized in Tab. 1. These baselines cover raw-logit, softmax-response, margin-based, OOD-style, and neighbor-based signals to evaluate the benefit of combining diverse confidence signals. Hyperparameter details remain in App. C. As a learned black-box baseline, AURC-MLP trains an MLP on fixed-classifier raw logits using SoftRank-AURC, with one hidden layer of width $2K$ for tasks with fewer than 1000 classes and two hidden layers of widths $2K$ and $K/2$ otherwise, where K is the logit input dimension.

Table 1: Overview of the ten post-hoc confidence-score baselines. Hyperparameter settings are provided in App. C.

Score	Signal family	Core definition or intuition
$RL_{\max}$ [27]	Raw logit	Maximum raw logit.
$SR_{\max}$ [4]	Softmax response	Maximum softmax probability.
SR_{doctor} [4]	Softmax response	Squared ℓ_2 norm of the softmax probability vector.
SR_{ent} [4]	Softmax response	Negative entropy of the softmax probability vector.
$RL_{\text{geo-}M}$ [4]	Margin	Signed distance to the closest decision hyperplane.
$RL_{\text{conf-}M}$ [4]	Margin	Gap between the two largest raw logits.
Energy [16]	Logit / OOD	Log-sum-exp aggregation of the raw logits.
KNN [28]	Neighbors / OOD	Distance-based confidence from nearest reference samples.
ViM [19]	Feature / OOD	Combines logits with the feature residual relative to a principal subspace.
SIRC [24]	Composite / OOD	Combines softmax confidence with an OOD-sensitive score.

Calibration protocol. For each image/NLP dataset or multiple-choice benchmark, we cache fixed-classifier logits, features, and correctness labels on a deployment-distribution evaluation pool. Unless otherwise stated, main results use $N = 1000$ calibration samples, 10% validation, and a disjoint 40% test split; only the calibration-size ablation varies N. We tune β on the validation split over $\{0.1, 0.5, 1, 2, 5, 10, 20\}$, and tune κ over the same grid for SoftRank-AURC. Learned models are optimized with PyTorch L-BFGS. For numerical stability, we add $\varepsilon_\alpha = 10^{-8}$ to the argument of the logarithm in $\widehat{\alpha}_i$. All reported AURC values are averaged over five different random seeds. The corresponding standard deviations are reported in App. C.

4.1 Image and NLP Datasets

Dataset setup. **Generalized-SC stress test.** We choose the image/NLP suite to stress-test post-hoc ensemble selection in the generalized SC setting in [4], where deployment errors may come from in-distribution mistakes, label-shift/OOD samples, and covariate-shifted inputs. Most datasets and fixed models follow this protocol, so the evaluation covers heterogeneous deployment pools where a single hand-crafted confidence score is unlikely to be uniformly reliable. **Dataset coverage.** The vision benchmarks include ImageNet [29], ImageNet-O, and OpenImage-O; CIFAR-10 [30], CIFAR-10-C, and CIFAR-100; iWildCam in-distribution and out-of-distribution splits [31]; Rendered-SST2 [32]; Caltech101 [33]; PatchCamelyon [34]; EyePACS; and NIH-CXR-LT, a long-tailed chest X-ray benchmark [35]. The NLP benchmarks include ADE-Corpus-V2, which is annotated for adverse drug events [36], and the Amazon Review dataset [37]. Together, these datasets cover clean in-distribution evaluation, synthetic corruptions, natural covariate shift, label shift, biomedical deployment, and long-tailed class imbalance. **Calibration details.** For the two small-pool datasets, Caltech101 and Rendered-SST2, we use 50% of the test pool for calibration because fewer than 1000 calibration samples are available under the standard split. Additional dataset and model details are provided in App. C.

Table 2: AURC performance ($\times 100$) across 15 datasets. **Blue**-background rows are our methods. **Green** text marks best or tied-best scores, and **Blue** text marks second-best or tied second-best scores, both at displayed precision. Abbreviations: AMZ = Amazon, CAL = Caltech101, C10/C1C/C100 = CIFAR-10/CIFAR-10-C/CIFAR-100, EYE = EyePACS, IMN/INO = ImageNet/ImageNet-O, WID/-WOD = iWildCam-ID/OOD, OIO = OpenImage-O, NIH = NIH-CXR-LT, PCM = PatchCamelyon, SST = Rendered-SST2.

Method	ADE	AMZ	CAL	C10	C1C	C100	EYE	IMN	INO	WID	WOD	OIO	NIH	PCM	SST
RL_{max}	87.53	12.55	2.21	1.64	7.49	28.44	11.68	4.88	5.94	23.91	28.61	12.02	77.05	15.70	13.11
SR_{max}	87.06	12.32	1.01	1.45	7.36	27.81	9.33	4.17	5.67	9.82	10.20	12.91	78.30	9.05	**9.07**
SR_{doctor}	85.39	12.27	1.01	1.45	7.36	27.81	9.36	4.23	5.71	9.92	10.22	12.86	77.65	9.05	**9.07**
SR_{ent}	85.25	**12.13**	0.98	1.47	7.37	27.87	9.44	5.43	6.85	10.34	10.66	13.96	77.61	9.05	**9.07**
ViM	73.62	14.40	4.06	1.06	11.09	36.71	10.27	8.66	9.51	16.66	12.13	18.86	75.39	11.22	14.66
$SIRC$	56.53	12.23	1.04	1.42	7.53	28.56	**9.11**	4.01	5.30	9.70	10.02	12.39	77.91	9.12	**9.06**
KNN	55.18	18.38	5.46	1.00	6.66	23.87	17.77	15.13	16.89	15.76	10.67	42.30	72.55	11.17	12.13
$Energy$	83.13	12.86	2.37	1.64	7.48	28.43	12.51	6.24	7.26	28.95	35.38	13.28	74.26	16.37	13.45
$RL_{\text{conf-M}}$	87.64	12.56	1.27	1.05	6.80	25.70	9.75	3.25	4.80	9.97	10.12	12.84	78.25	9.05	**9.07**
$RL_{\text{geo-M}}$	88.20	12.87	1.27	1.10	7.21	25.75	15.52	3.27	4.78	9.97	10.12	12.77	77.73	9.05	**9.07**
AURC-MLP	55.82	16.92	3.41	2.27	7.14	24.18	13.20	10.18	11.62	**7.58**	**8.28**	15.23	**71.12**	**3.72**	18.55
SELE-SCR	55.50	12.15	0.75	**0.98**	6.54	24.07	9.20	**3.22**	4.59	9.43	8.57	11.28	72.47	8.43	9.11
SELE-SLR	**53.16**	**12.13**	**0.72**	**0.98**	**6.53**	23.91	**9.19**	**3.22**	**4.58**	9.40	8.49	**11.20**	72.39	6.46	9.10
AURC-SCR	55.25	12.14	0.73	**0.97**	6.54	**23.79**	9.21	**3.24**	**4.58**	9.32	8.36	11.23	72.45	8.33	9.08
AURC-SLR	**53.21**	**12.12**	**0.68**	**0.97**	**6.48**	**23.76**	9.23	3.25	**4.56**	**9.07**	**8.16**	**11.07**	**72.25**	**6.45**	**9.07**

Results. **(1) Learned ensembles are consistently competitive.** In Tab. 2, the learned ensemble rows attain the best or tied-best reported AURC on 10 of the 15 image/NLP datasets: ADE-Corpus-V2, Amazon, Caltech101, CIFAR-10, CIFAR-10-C, CIFAR-100, ImageNet, ImageNet-O, iWildCam-OOD, and OpenImage-O. This pattern is consistent with our premise that selective-classification performance varies across deployment pools and can benefit from learning how to combine multiple confidence signals. **(2) SoftRank-AURC and SELE both provide strong AURC-oriented training.** SELE is also an AURC-oriented surrogate, so the comparison should be read as SoftRank-AURC plug-in training versus SELE pairwise training rather than AURC-oriented versus non-AURC-oriented training. SoftRank-AURC-SLR attains the best or tied-best value among all methods on Amazon, Caltech101, CIFAR-10, CIFAR-10-C, CIFAR-100, ImageNet-O, iWildCam-OOD, and OpenImage-O, while SELE-SLR or SELE-SCR attains the best or tied-best value on ADE-Corpus-V2 and ImageNet. Thus the table favors SoftRank-AURC on more columns, but it also shows that SELE remains a strong AURC-oriented baseline. **(3) Ensembling improves robustness when single scores disagree, with clear exceptions.** The ensemble advantage is visible on shifted or otherwise challenging deployment pools such as CIFAR-10-C, ImageNet-O, iWildCam-OOD, and OpenImage-O. The exceptions are EyePACS and Rendered-SST2, where SIRC is best, and iWildCam-ID, NIH-CXR-LT, and PatchCamelyon, where AURC-MLP is best. These exceptions reinforce that the best confidence signal depends on the dataset, model, and deployment distribution; the learned ensembles provide a compact way to adapt this choice from calibration data.

4.2 Multiple-Choice Benchmarks

Benchmark setup. **Classification interface.** We use LLM multiple-choice benchmarks to test post-hoc ensemble selection when the fixed classifier is a generative language model and confidence must be extracted from answer-option logits. The multiple-choice format gives a clean classification problem: the label set is the candidate options, the model is not retrained, and all confidence scores are computed from comparable candidate logits. **Task and model coverage.** We evaluate MMLU [38], MMLU-Pro [39], HellaSwag [40], and ARC-Easy from the AI2 Reasoning Challenge [41], covering broad knowledge, harder multi-domain reasoning, commonsense completion, and science reasoning. We use Mistral-7B-v0.1 [42] and Qwen2.5-7B [43] as representative open-weight 7B language models from different model families. **Candidate-logit scoring.** For each question, we restrict computation to the K candidate option-token logits. Because LLM logits can have large sample-dependent magnitudes, directly applying softmax can make one option dominate the probability vector; following the scale-sensitivity observation for softmax-response scores in [4], we divide each sample's candidate-logit vector by its mean logit magnitude before computing confidence scores. This positive sample-wise scaling preserves the argmax prediction while preventing softmax-based

Table 3: AURC performance ($\times 100$) on four LLM benchmarks with calibration size $N = 1000$. **Blue**-background rows are our methods. **Green** text marks best or tied-best scores, and **Blue** text marks second-best or tied second-best scores, both at displayed precision.

Method	MMLU Mistral	MMLU Qwen	MMLU-Pro Mistral	MMLU-Pro Qwen	HellaSwag Mistral	HellaSwag Qwen	ARC-Easy Mistral	ARC-Easy Qwen
$RL_{\max}$	18.56	10.53	51.91	34.38	37.85	2.48	2.66	0.37
$SR_{\max}$	18.54	10.48	51.88	34.27	**37.83**	2.46	2.63	0.37
SR_{doctor}	19.17	11.76	53.06	38.33	39.10	3.13	3.12	0.57
SR_{ent}	19.22	11.86	53.17	38.80	39.13	3.18	3.19	0.60
ViM	56.11	40.10	78.00	68.65	63.20	16.15	13.11	3.10
$SIRC$	32.10	20.73	69.71	63.98	57.69	7.17	4.86	0.98
KNN	55.85	42.64	78.46	69.08	63.34	16.95	13.12	2.04
$Energy$	19.29	11.97	53.29	39.35	39.16	3.24	3.27	0.62
$RL_{\text{conf-M}}$	18.66	10.32	**51.87**	34.30	38.78	2.28	2.48	**0.32**
$RL_{\text{geo-M}}$	18.76	11.21	56.47	42.10	45.26	2.34	2.50	**0.33**
AURC-MLP	21.16	15.77	56.14	44.28	43.47	5.08	5.95	1.83
SELE-SCR	**18.30**	**10.28**	51.99	33.68	**37.83**	**2.27**	**2.44**	**0.32**
SELE-SLR	**18.31**	**10.28**	51.98	**33.65**	37.84	**2.27**	**2.44**	**0.32**
AURC-SCR	18.32	**10.26**	51.99	**33.62**	**37.82**	**2.26**	**2.45**	**0.32**
AURC-SLR	18.33	10.30	**51.79**	**33.65**	**37.83**	**2.26**	**2.45**	**0.32**

scores from being driven mainly by raw logit scale. Additional LLM dataset and model details are provided in App. C.

Results. **(1) Broad transfer to generative-model classifiers.** In Tab. 3, learned ensembles attain the best or tied-best reported AURC on all eight benchmark-backbone pairs: MMLU-Mistral, MMLU-Qwen, MMLU-Pro-Mistral, MMLU-Pro-Qwen, HellaSwag-Mistral, HellaSwag-Qwen, ARC-Easy-Mistral, and ARC-Easy-Qwen. On ARC-Easy-Qwen, $RL_{\text{conf-M}}$, SELE-SCR, SELE-SLR, SoftRank-AURC-SCR, and SoftRank-AURC-SLR all round to 0.32, so this column should be read as a tie at the reported precision. **(2) Objective comparison.** Across columns that distinguish the learned objectives at displayed precision, SoftRank-AURC-SCR or SoftRank-AURC-SLR attains the best or tied-best learned-objective value on MMLU-Qwen, MMLU-Pro-Mistral, MMLU-Pro-Qwen, HellaSwag-Mistral, and HellaSwag-Qwen, while SELE-SCR or SELE-SLR attains the best or tied-best learned-objective value on MMLU-Mistral and ARC-Easy-Mistral; ARC-Easy-Qwen does not distinguish the learned objectives at the displayed precision. Thus the LLM table favors SoftRank-AURC on more columns that distinguish the learned objectives, but it does not diminish SELE: SELE attains or shares the best learned-objective value on MMLU-Mistral, ARC-Easy-Mistral, and ARC-Easy-Qwen. **(3) Why the margins are small.** After logit rescaling, raw-logit, softmax-response, and margin-based scores are already strong for multiple-choice prediction, so the absolute gains are typically modest. The learned SCR/SLR selectors nevertheless avoid the poor ViM, KNN, SIRC, and AURC-MLP results on these LLM tasks, suggesting that deployment-calibrated combinations are useful even when several simple logit-derived scores are competitive.

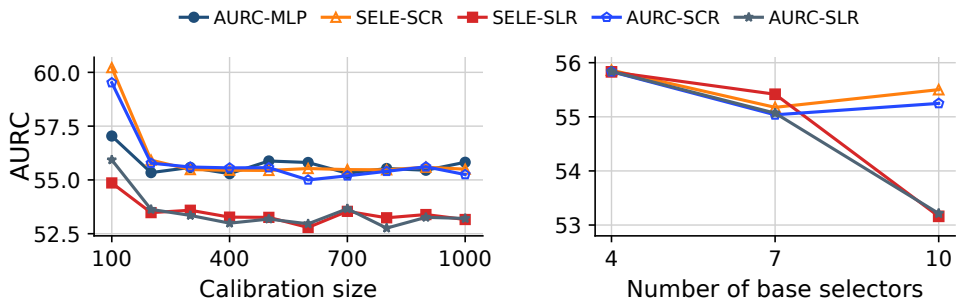

Figure 1: ADE-Corpus-V2 ablation studies. The left side shows the calibration-size ablation for the four main ensemble methods and the AURC-MLP baseline. The right side shows the ablation study of the number of base selectors at calibration size $N = 1000$; lower is better.

4.3 Ablation Studies

We report two ADE-Corpus-V2 ablations in Fig. 1. **(1) Calibration size.** We vary the calibration size N while keeping validation and test splits fixed, and compare the four main ensemble methods with the AURC-MLP baseline. The learned ensembles improve after the smallest calibration split and then flatten, consistent with the fact that these selectors have relatively few trainable parameters and can estimate useful weights and thresholds from modest labeled calibration sets. The SLR variants remain particularly strong across the range, suggesting that the thresholded-selector parameterization is stable in the small-calibration regime. **(2) Number of base selectors.** We vary the number of base selectors at $N = 1000$. On ADE-Corpus-V2, the SLR variants benefit from the full selector set: each base score is compared with its own learned threshold before combination, so the model can use whether a score provides a useful selector signal rather than relying only on the exact raw magnitude of heterogeneous score values. Additional scores can therefore act as complementary thresholded signals. Reduced sets remain competitive, suggesting limited dependence on any single large score pool. We report the full results for both ablations in Appendix C, where we also examine how the ensembles weight each base scorer.

5 Conclusion

We proposed an ensemble selective-classification framework for the post-hoc setting that learns to combine multiple heterogeneous confidence scores using a labeled calibration set while keeping the base classifier fixed. The framework can be trained with either our SoftRank-AURC plug-in objective or the pairwise SELE objective, giving two AURC-oriented ways to learn the same score-combiner and selector-combiner families. Across the 15 image and NLP datasets, learned ensembles attain the best or tied-best reported AURC on 10 datasets; SoftRank-AURC-SLR is best or tied-best on eight of these datasets, while SELE variants are best or tied-best on ADE-Corpus-V2 and ImageNet. On the LLM multiple-choice benchmarks, learned ensembles attain the best or tied-best reported AURC on all eight benchmark-backbone pairs, with SoftRank-AURC leading more columns that distinguish the learned objectives and SELE attaining or sharing the best learned-objective value on MMLU-Mistral, ARC-Easy-Mistral, and ARC-Easy-Qwen. These results suggest that deployment-calibrated ensembling is a simple way to adapt among heterogeneous post-hoc confidence signals and improve selective classification without retraining the base classifier.

6 Limitations

Our framework is post-hoc: it learns a better confidence ranking but does not change the fixed base classifier, so it cannot correct systematic errors unless the available confidence scores expose useful signals for them. The method also relies on a labeled calibration set from the deployment distribution; if that distribution shifts, the learned weights and thresholds may need to be revalidated or retuned. Finally, SoftRank-AURC uses pairwise soft comparisons and therefore has quadratic cost in the calibration size, which is manageable in our experiments but may require approximations for larger deployments.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Institutes of Health (NIH) under award No. R01CA287413, R01NS131314, R01MH138929, U01FD008720 and by the National Science Foundation (NSF) under award No. 2435911. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors acknowledge the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing resources that contributed to the research results reported within this paper. The authors declare no competing interests.

References

[1] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[2] V. Franc, D. Prusa, and V. Voracek, "Optimal strategies for reject option classifiers," *Journal of Machine Learning Research*, vol. 24, no. 11, pp. 1–49, 2023.
[3] A. Pugnana, L. Perini, J. Davis, and S. Ruggieri, "Deep neural network benchmarks for selective classification," *Journal of Data-centric Machine Learning Research*, 2024. [Online]. Available: https://openreview.net/forum?id=xDPzHbtAEs
[4] H. Liang, L. Peng, and J. Sun, "Selective classification under distribution shifts," *Transactions on Machine Learning Research*, vol. 2024, 2024.
[5] G. Luo, Y. Wu, S. Zhang, A. Deshwal, and J. Sun, "Aligning language models with selective prediction," *arXiv preprint arXiv:2607.03528*, 2026.
[6] A. Heng and H. Soh, "Know when to abstain: Optimal selective classification with likelihood ratios," *arXiv preprint arXiv:2505.15008*, 2025.
[7] ARPA-H, "Precise-ai: Performance and reliability evaluation for continuous modifications and useability of artificial intelligence," https://arpa-h.gov/explore-funding/programs/precise-ai, 2025, accessed: October 1, 2025.
[8] H. Yao, C. Choi, B. Cao, Y. Lee, P. W. W. Koh, and C. Finn, "Wild-time: A benchmark of in-the-wild distribution shift over time," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 309–10 324, 2022.
[9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY: Springer, 2009.
[10] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. CRC Press, 2025.
[11] C. Chow, "On optimum recognition error and reject tradeoff," *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 41–46, 1970.
[12] R. El-Yaniv and Y. Wiener, "On the foundations of noise-free selective classification," *Journal of Machine Learning Research*, vol. 11, no. 53, pp. 1605–1641, 2010. [Online]. Available: http://jmlr.org/papers/v11/el-yaniv10a.html
[13] Y. Geifman, G. Uziel, and R. El-Yaniv, "Bias-reduced uncertainty estimation for deep neural classifiers," *arXiv preprint arXiv:1805.08206*, 2018.
[14] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.
[15] L. Feng, M. O. Ahmed, H. Hajimirsadeghi, and A. Abdi, "Towards better selective classification," *arXiv preprint arXiv:2206.09034*, 2022.
[16] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 464–21 475, 2020.
[17] S. Fort, J. Ren, and B. Lakshminarayanan, "Exploring the limits of out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 7068–7081, 2021.
[18] Y. Sun, C. Guo, and Y. Li, "React: Out-of-distribution detection with rectified activations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 144–157, 2021.
[19] H. Wang, Z. Li, L. Feng, and W. Zhang, "Vim: Out-of-distribution with virtual-logit matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4921–4930.
[20] H. Zhou, J. Van Landeghem, T. Popordanoska, and M. B. Blaschko, "A novel characterization of the population area under the risk coverage curve (aurc) and rates of finite sample estimators," *arXiv preprint arXiv:2410.15361*, 2024.
[21] H. Zhou, S. G. Gruber, T. Popordanoska, and M. B. Blaschko, "Revisiting reweighted risk for calibration: Aurc, focal, and inverse focal loss," *arXiv preprint arXiv:2505.23463*, 2025.
[22] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[23] Q. Ding, Y. Cao, and P. Luo, "Top-ambiguity samples matter: Understanding why deep ensemble works in selective classification," *Advances in Neural Information Processing Systems*, vol. 36, pp. 35 497–35 521, 2023.
[24] G. Xia and C.-S. Bouganis, "Augmenting softmax information for selective classification with out-of-distribution data," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 1995–2012.
[25] J. Kim, J. Koo, and S. Hwang, "A unified benchmark for the unknown detection capability of deep neural networks," *Expert Systems with Applications*, vol. 229, p. 120461, 2023.
[26] R. E. Schapire and Y. Freund, *Boosting: Foundations and Algorithms*. MIT Press, 2012.
[27] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, and D. Song, "Scaling out-of-distribution detection for real-world settings," *arXiv preprint arXiv:1911.11132*,

2019.
[28] Y. Sun, Y. Ming, X. Zhu, and Y. Li, "Out-of-distribution detection with deep nearest neighbors," in *International Conference on Machine Learning*. PMLR, 2022, pp. 20 827–20 840.
[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vision*, vol. 115, no. 3, p. 211–252, Dec. 2015. [Online]. Available: https://doi.org/10.1007/s11263-015-0816-y
[30] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," University of Toronto, Toronto, Ontario, Tech. Rep. 0, 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf
[31] S. Beery, A. Agarwal, E. Cole, and V. Birodkar, "The iwildcam 2021 competition dataset," *arXiv preprint arXiv:2105.03494*, 2021.
[32] 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 *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html
[33] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *2004 Conference on Computer Vision and Pattern Recognition Workshop*. IEEE, 2004, pp. 178–178.
[34] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Rotation equivariant cnns for digital pathology," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 210–218.
[35] G. Holste, S. Wang, Z. Jiang, T. C. Shen, G. Shih, R. M. Summers, Y. Peng, and Z. Wang, "Long-tailed classification of thorax diseases on chest x-ray: A new benchmark study," in *MICCAI Workshop on Data Augmentation, Labelling, and Imperfections*. Springer, 2022, pp. 22–32.
[36] H. Gurulingappa, A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, and L. Toldo, "Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports," *Journal of Biomedical Informatics*, vol. 45, no. 5, pp. 885–892, 2012.
[37] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 188–197. [Online]. Available: https://aclanthology.org/D19-1018/
[38] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2020.
[39] 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 *Proceedings of the 38th International Conference on Neural Information Processing Systems*, ser. NIPS '24. Red Hook, NY, USA: Curran Associates Inc., 2024.
[40] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "Hellaswag: Can a machine really finish your sentence?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4791–4800.
[41] 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 arXiv:1803.05457*, 2018.
[42] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023. [Online]. Available: https://arxiv.org/abs/2310.06825
[43] Q. Team, "Qwen2.5: A party of foundation models," September 2024. [Online]. Available: https://qwenlm.github.io/blog/qwen2.5/
[44] T. M. Apostol, *Mathematical Analysis*, 2nd ed. Pearson, 1974.

[45] N. Falkner and G. Teschl, "On the substitution rule for lebesgue–stieltjes integrals," *Expositiones Mathematicae*, vol. 30, no. 4, pp. 412–418, 2012.
[46] P. Billingsley, "Probability and measure. 3rd wiley," *New York*, 1995.
[47] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "Spanbert: Improving pre-training by representing and predicting spans," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64–77, 2020.
[48] H. Yao, Y. Wang, S. Li, L. Zhang, W. Liang, J. Zou, and C. Finn, "Improving out-of-distribution robustness via selective augmentation," in *Proceedings of the Thirty-ninth International Conference on Machine Learning*, 2022.

A Technical Details of the Proposed Method

A.1 Scorer Normalization

Without loss of generality, we assume that every base scorer is monotonically increasing in the confidence level, i.e., the higher the confidence level, the higher the score value; we can simply negate the score values if any scorer is monotonically decreasing. Since these scorers may assign confidence values of different scales and distributions, we need to normalize them to a common scale and similar distributions before combining them. To normalize each scorer, it is tempting to consider the

Table 4: Skewness and kurtosis of the score distribution induced by each base scorer, averaged over the 15 image and NLP datasets we use in our experiments of Sec. 4. A Gaussian has skewness 0 and kurtosis 3. The skewness measures the asymmetry of a distribution, and the kurtosis measures the heaviness of its tails (a Gaussian has kurtosis 3, and larger values indicate heavier tails). Bold marks the three most skewed and heavy-tailed scores.

Score	$\mathrm{RL}_{\max}$	$\mathrm{SR}_{\max}$	$\mathrm{SR}_{\mathrm{doctor}}$	$\mathrm{SR}_{\mathrm{ent}}$	ViM	SIRC	KNN	Energy	$\mathrm{RL}_{\text{conf-M}}$	$\mathrm{RL}_{\text{geo-M}}$
Skewness	-0.33	$\mathbf{-17.74}$	$\mathbf{-6.20}$	$\mathbf{-7.30}$	-0.34	-1.31	-1.55	-0.23	-1.20	-1.20
Kurtosis	2.7	**386.3**	**47.3**	**73.7**	2.9	4.6	6.3	2.4	4.2	4.2

standard zero-mean-unit-variance normalization. However, both mean and variance are sensitive to outlying probability masses, and they could be ill-defined for certain heavy-tailed distributions. These issues indeed occur for our score distributions: as shown in Tab. 4, most base-score distributions are substantially skewed and heavy-tailed, and thus the standard normalization might be ill here. We therefore use a robust alternative based on the median and the median absolute deviation (MAD). For each $t \in [T]$, the *median* m_t and *median absolute deviation* MAD_t are

$$m_t \doteq \mathrm{median}\left\{\tilde{s}_t(\boldsymbol{x}_i) : i \in [N]\right\}, \qquad \mathrm{MAD}_t \doteq \mathrm{median}\left\{|\tilde{s}_t(\boldsymbol{x}_i) - m_t| : i \in [N]\right\}.$$

With these, we define the median-MAD (MM) normalization as

$$\tilde{s}_t^{\mathrm{MM}}(\boldsymbol{x}) = (\tilde{s}_t(\boldsymbol{x}) - m_t)/\mathrm{MAD}_t.$$

But this still leaves outliers as outliers, and different scorers may still have different scales. So, we further squash the MM-normalized scorers into a common output range $(0, 1)$ using a sigmoid transformation, which meanwhile suppresses potential outliers, yielding the final normalized scorer

$$s_t(\boldsymbol{x}) = \sigma(\tilde{s}_t^{\mathrm{MM}}(\boldsymbol{x})/c_{\mathrm{MAD}}) \in (0, 1), \quad \text{with } \sigma(u) \doteq \left(1 + e^{-u}\right)^{-1}$$

where $c_{\mathrm{MAD}} \doteq 1/\Phi^{-1}(3/4) \approx 1.4826$ is a fixed constant and Φ is the CDF of the standard normal distribution. So, our complete transformation reads

$$s_t(\boldsymbol{x}) = \sigma\left((\tilde{s}_t(\boldsymbol{x}) - m_t)/b_t\right) \in (0, 1), \quad \text{with } b_t \doteq \max\left(c_{\mathrm{MAD}} \cdot \mathrm{MAD}_t,\ \delta_{\min}\right), \tag{A.1}$$

where $\delta_{\min} = 10^{-3}$ is to prevent excessively small scale in $c_{\mathrm{MAD}} \cdot \mathrm{MAD}_t$. Note that this transformation is monotonically increasing in $\tilde{s}_t(\boldsymbol{x})$, so it preserves the confidence ranking across samples.

A.2 An Alternative Proof of the Population AURC Formula

Here, we provide a self-contained, first-principle derivation of the population AURC formula used in the design of our finite-sample SoftRank-AURC loss. The formula first appears in [20], but their derivation is very involved with substantial missing details.

We start with reminding a few basic facts about Riemann-Stieltjes integrals and expectations, which will be used in our derivation.

Theorem 3 (Integration by parts for Riemann-Stieltjes integrals; Theorem 7.6 of [44]). *Let $a < b$ and consider $f, g : [a, b] \to \mathbb{R}$. Let df and dg denote differentials in the Riemann-Stieltjes sense. Then, one of $\int_a^b f(x)\, dg(x)$ and $\int_a^b g(x)\, df(x)$ is defined if the other is defined, and they are connected by the following integration by parts formula:*

$$\int_a^b f(x)\, dg(x) + \int_a^b g(x)\, df(x) = f(b)g(b) - f(a)g(a). \tag{A.2}$$

Theorem 4 (Change-of-variable formula for Lebesgue-Stieltjes integrals; Eqs. (1)–(2) of [45]). *Let $a < b$ and let $h : [a, b] \to \mathbb{R}$ be a continuous non-decreasing function. Define the generalized inverse (quantile function) $h^{-1} : [h(a), h(b)] \to [a, b]$ by $h^{-1}(u) \doteq \inf\{x \in [a, b] : h(x) \geq u\}$. Consider $f, g : [h(a), h(b)] \to \mathbb{R}$, where f is continuous and non-decreasing, and g is Borel measurable. Assume that $\int_{h(a)}^{h(b)} g(y)\, df(y)$ exists in the Lebesgue–Stieltjes sense. We have*

$$\int_{h(a)}^{h(b)} g(y)\, df(y) = \int_a^b [g \circ h](x)\, d[f \circ h](x). \tag{A.3}$$

or equivalently, expressed via the generalized inverse:

$$\int_a^b [g \circ h](x)\, d[f \circ h](x) = \int_{h(a)}^{h(b)} [g \circ h]\left(h^{-1}(u)\right)\, df(u). \tag{A.4}$$

Note that we set $N \doteq f$, $M \doteq h$, and $f \doteq g \circ h$ in their original statement (red indicates their original symbols) in [45].

Theorem 5 (Change-of-variable formula for expectations; Theorem 16.13 of [46]). *Consider two measurable spaces (X_1, Σ_1) and (X_2, Σ_2). Let X be a random variable taking values in X_1 with distribution P. Let $f : X_1 \to X_2$ be a measurable function, and define the random variable $Y = f(X)$ taking values in X_2. For any measurable function $g : X_2 \to \mathbb{R}$, $g(Y)$ is integrable if and only if $g(f(X))$ is integrable. In this case, we have:*

$$\mathbb{E}_Y[g(Y)] = \mathbb{E}_X[g(f(X))]. \tag{A.5}$$

Our much simplified derivation of the popular AURC formula starts with the canonical definition of the AURC: $\mathrm{AURC}(f, s) = \int_0^1 R_{s,\gamma}(\phi_{s,\gamma})\, d\phi_{s,\gamma}$.

Proposition 6 (Population AURC as weighted risk). *Consider any data distribution $\mathcal{D}$ on $\mathcal{X} \times \{1, -1\}$ with marginal distributions $\mathcal{D}_{\boldsymbol{x}}$ and $\mathcal{D}_y$, respectively. Let $s : \mathcal{X} \to \mathbb{R}$ be a measurable scorer such that $s(\boldsymbol{x})$ has a continuous CDF. We have the following formula for the population AURC:*

$$\mathrm{AURC}(f, s) = \mathbb{E}_{\boldsymbol{x},y}[-\log(1 - \mathbb{E}_{\boldsymbol{x}' \sim \mathcal{D}_{\boldsymbol{x}}} \mathbb{I}\{s(\boldsymbol{x}') \leq s(\boldsymbol{x})\}) \cdot \mathbb{I}\{y = -1\}]. \tag{A.6}$$

Proof. If $\Pr(y = -1) = 0$, both sides are zero; hence assume $\Pr(y = -1) > 0$. First, by viewing $\phi_{s,\gamma}$ as the integration variable, we have

$$1 \geq \int_0^1 R_{s,\gamma}(\phi_{s,\gamma})\, d\phi_{s,\gamma} = \int_0^1 \frac{E(\phi_{s,\gamma})}{\phi_{s,\gamma}}\, d\phi_{s,\gamma} = \int_0^1 E(\phi_{s,\gamma})\, d\log\phi_{s,\gamma} \geq 0, \tag{A.7}$$

where $E(\phi_{s,\gamma}) \doteq \mathbb{E}_{\boldsymbol{x},y}[\mathbb{I}\{s(\boldsymbol{x}) \geq \gamma\}\mathbb{I}\{y = -1\}]$. Write $\phi(\gamma) \doteq \phi_{s,\gamma} = \Pr(s(\boldsymbol{x}) \geq \gamma)$. The continuous CDF assumption makes $u \mapsto \phi(-u)$ continuous and nondecreasing. Applying Thm. 4 to this map and then substituting $\gamma = -u$ gives the reversed limits below. Integrals involving $\log\phi$ are understood as limits over positive coverage. The boundary terms in integration by parts vanish since $0 \leq E(c) \leq c$, $c\log c \to 0$ as $c \downarrow 0$, and $\log 1 = 0$.

$$\int_\infty^{-\infty} E(\phi(\gamma))\, d\log\phi(\gamma) \tag{A.8}$$

$$= \int_{-\infty}^{\infty} \log\phi(\gamma)\, dE(\phi(\gamma)) \quad \text{(by Thm. 3)} \tag{A.9}$$

$$= \int_{-\infty}^{\infty} \log \phi(\gamma)\, d\mathbb{E}_{\boldsymbol{x},y}[\mathbb{I}\{s(\boldsymbol{x}) \geq \gamma\}\, \mathbb{I}\{y = -1\}] \tag{A.10}$$

$$= \mathbb{P}^{-} \int_{-\infty}^{\infty} \log \phi(\gamma)\, d\mathbb{E}_{\boldsymbol{x}|y=-1}[\mathbb{I}\{s(\boldsymbol{x}) \geq \gamma\}] \quad (\mathbb{P}^{-} \doteq \mathbb{P}\{y = -1\}) \tag{A.11}$$

$$= \mathbb{P}^{-} \int_{-\infty}^{\infty} \log \phi(\gamma)\, d(1 - \mathrm{CDF}_{s(\boldsymbol{x})|y=-1}(\gamma)) \quad \text{(by Thm. 5)} \tag{A.12}$$

$$= \mathbb{P}^{-} \int_{-\infty}^{\infty} -\log \phi(\gamma)\, d\,\mathrm{CDF}_{s(\boldsymbol{x})|y=-1}(\gamma) \tag{A.13}$$

$$= \mathbb{P}^{-} \mathbb{E}_{s(\boldsymbol{x})|y=-1}[-\log \phi(s(\boldsymbol{x}))] \quad \text{(Lebesgue form of expectation)} \tag{A.14}$$

$$= \mathbb{E}_{s(\boldsymbol{x}),y}[-\log \phi(s(\boldsymbol{x}))\mathbb{I}\{y = -1\}] \tag{A.15}$$

$$= \mathbb{E}_{\boldsymbol{x},y}[-\log \phi(s(\boldsymbol{x}))\mathbb{I}\{y = -1\}] \quad \text{(by Thm. 5)} \tag{A.16}$$

$$= \mathbb{E}_{\boldsymbol{x},y}[-\log(1 - \mathbb{E}_{\boldsymbol{x}' \sim \mathcal{D}_{\boldsymbol{x}}} \mathbb{I}\{s(\boldsymbol{x}') \leq s(\boldsymbol{x})\}) \cdot \mathbb{I}\{y = -1\}], \quad \text{(by Thm. 5)} \tag{A.17}$$

completing the proof. □

A.3 The Unified Training Algorithm

Algorithm 3 Unified Ensemble Training with Alternative Objectives

Input: Raw post-hoc scorers $\{\tilde{s}_t\}_{t=1}^{T}$; calibration set Z_B^{cali} with labels $\{y_i^z\}$; ensemble form $\mathcal{F} \in \{S_{\mathrm{scr}}, S_{\mathrm{slr}}^{\beta}\}$; loss type $\mathcal{L} \in \{\text{SoftRank-AURC}, \text{SELE}\}$; smoothing temperature $\beta > 0$ for S_{slr}^{β}; SoftRank-AURC temperature $\kappa > 0$ if applicable; optimizer; outer iterations E.

Output: Learned weights $\boldsymbol{\omega}^*$; inner thresholds $\boldsymbol{\gamma}^*$ when $\mathcal{F} = S_{\mathrm{slr}}^{\beta}$.

Normalize scores: Apply the score-normalization procedure in App. A.1 to obtain normalized score values $\{s_t(\boldsymbol{x}_i) : t \in [T], i \in [N]\}$.
Initialize: Unconstrained weight parameter vector $\boldsymbol{\theta}_{\boldsymbol{\omega}}$.
if $\mathcal{F} = S_{\mathrm{slr}}^{\beta}$ **then**
    Initialize unconstrained threshold parameter vector $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$.
end if
for epoch $= 1$ to E **do**
    Compute ensemble weights: $\boldsymbol{\omega} \leftarrow \mathrm{abs}(\boldsymbol{\theta}_{\boldsymbol{\omega}})$.
    if $\mathcal{F} = S_{\mathrm{slr}}^{\beta}$ **then**
        Compute inner thresholds: $\boldsymbol{\gamma} \leftarrow \sigma(\boldsymbol{\theta}_{\boldsymbol{\gamma}})$ elementwise.
    end if
    Compute ensemble scores $S(\boldsymbol{x}_i)$ using the chosen form $\mathcal{F}$ via Eqs. (3.4) and (3.9).
    if $\mathcal{L} =$ SoftRank-AURC **then**
        Compute SoftRank-AURC loss via Eq. (3.10).
        Update the trainable ensemble parameters using the optimizer on $\mathcal{L}_A$.
    else
        Compute SELE loss via Eq. (3.11).
        Update the trainable ensemble parameters using the optimizer on $\mathcal{L}_S$.
    end if
end for
Return: $\boldsymbol{\omega}^* = \mathrm{abs}(\boldsymbol{\theta}_{\boldsymbol{\omega}})$, and $\boldsymbol{\gamma}^* = \sigma(\boldsymbol{\theta}_{\boldsymbol{\gamma}})$ with σ applied elementwise if $\mathcal{F} = S_{\mathrm{slr}}^{\beta}$.

B Proofs for Theoretical Results

Throughout these proofs, the normalized score maps $\{s_t\}_{t=1}^{T}$ are treated as fixed measurable functions, exactly as in the theoretical idealization in Sec. 3.3. All randomness is therefore with respect to the i.i.d. calibration sample, the induced binary labels y^z, and any fresh deployment draw.

In the main text and algorithms, $s_t : \mathcal{X} \to (0, 1)$ denotes the normalized score in Eq. (A.1). In the theory, we use the fixed-preprocessing idealization: the normalization statistics are treated as fixed, for example because they are computed on an independent preprocessing split or replaced

by population analogues, and all randomness comes from the calibration sample used to learn the ensemble. Equivalently, the analysis is conditional on the preprocessing step and focuses on the statistical complexity of learning the ensemble selector once the normalized base scores are fixed.

B.1 Normalization Properties

Lemma 7 (Normalization properties)**.** *For every $\boldsymbol{\omega} \in \Delta^T$, $\boldsymbol{\gamma} \in [0,1]^T$, and $\beta > 0$, the selector $S_{\boldsymbol{\omega},\boldsymbol{\gamma}}$ in Eq. (3.9) satisfies:*

(i) ***Monotonicity.*** *Each $s_t(\cdot)$ is strictly increasing in $\tilde{s}_t(\cdot)$.*

(ii) ***Boundedness.*** *For all $\boldsymbol{x}$, $s_t(\boldsymbol{x}) \in (0,1)$ and hence $\tanh((s_t(\boldsymbol{x}) - \gamma_t)/\beta) \in (-1,1)$; thus $S_{\boldsymbol{\omega},\boldsymbol{\gamma}}(\boldsymbol{x}) \in (-1,1)$.*

(iii) ***Lipschitzness.*** *For any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, $|S_{\boldsymbol{\omega},\boldsymbol{\gamma}}(\boldsymbol{x}) - S_{\boldsymbol{\omega},\boldsymbol{\gamma}}(\boldsymbol{x}')| \le \frac{1}{\beta} \max_{t\in[T]} |s_t(\boldsymbol{x}) - s_t(\boldsymbol{x}')|$.*

Proof. For the normalized score construction in Eq. (A.1), the map $\tilde{s}_t \mapsto \sigma((\tilde{s}_t - m_t)/b_t)$ is the composition of an affine map with positive slope $1/b_t$ and the strictly increasing logistic function σ, so each $s_t(\cdot)$ is strictly increasing in the raw score $\tilde{s}_t(\cdot)$. This proves monotonicity.

Because σ takes values in $(0,1)$, each normalized score satisfies $s_t(\boldsymbol{x}) \in (0,1)$. Hence $\tanh((s_t(\boldsymbol{x}) - \gamma_t)/\beta) \in (-1,1)$ for every t, and since $\boldsymbol{\omega} \in \Delta^T$ is a nonnegative vector, $S_{\boldsymbol{\omega},\boldsymbol{\gamma}}(\boldsymbol{x}) \in (-1,1)$. This proves boundedness.

Finally, using $|\tanh'(u)| \le 1$, the mean-value theorem, and $\sum_{t=1}^T \omega_t \le 1$,

$$\begin{aligned} |S_{\boldsymbol{\omega},\boldsymbol{\gamma}}(\boldsymbol{x}) - S_{\boldsymbol{\omega},\boldsymbol{\gamma}}(\boldsymbol{x}')| &\le \sum\nolimits_{t=1}^T \omega_t \left| \tanh\left(\frac{s_t(\boldsymbol{x}) - \gamma_t}{\beta}\right) - \tanh\left(\frac{s_t(\boldsymbol{x}') - \gamma_t}{\beta}\right) \right| \\ &\le \frac{1}{\beta} \sum\nolimits_{t=1}^T \omega_t \, |s_t(\boldsymbol{x}) - s_t(\boldsymbol{x}')| \le \frac{1}{\beta} \max_{t\in[T]} |s_t(\boldsymbol{x}) - s_t(\boldsymbol{x}')|. \end{aligned}$$

This proves Lipschitzness. □

B.2 SoftRank-AURC oracle inequality

The proof of Theorem 2 combines two bounds. Lemma 9 controls the error from smoothing the population AURC, and Lem. 10 controls the sampling error of the smoothed loss uniformly over the ensemble scores. With the classifier f fixed, write $\mathcal{L}_A$ for the empirical SoftRank-AURC loss and $\mathcal{R}_A$ for its smoothed population counterpart. Their upper-tail formulas are given in Def. 8.

B.2.1 Proof of Theorem 2 (oracle inequality)

Proof. Let $\widehat{S}$ be an empirical minimizer. The case $N = 1$ follows from $0 \le \mathrm{AURC} \le 1$ after enlarging the absolute constant. For $N \ge 2$, work on the event of Lem. 10, and abbreviate its deviation bound by $\mathfrak{E}^{\mathrm{sr}}$. Let S be a comparator in the uniform closure $\overline{\mathcal{S}}$ of $\mathcal{S}$ such that the CDFs of both S and $\widehat{S}$ have Lipschitz constant at most L_G. By Lem. 9, both smoothing errors are at most B_κ. Empirical optimality gives $\mathcal{L}_A(\widehat{S}) \le \mathcal{L}_A(S)$, which extends to the closure by continuity. Combining these bounds gives

$$\begin{aligned} \mathrm{AURC}(\widehat{S}) &\le \mathcal{R}_A(\widehat{S}) + B_\kappa \le \mathcal{L}_A(\widehat{S}) + \mathfrak{E}^{\mathrm{sr}} + B_\kappa \\ &\le \mathcal{L}_A(S) + \mathfrak{E}^{\mathrm{sr}} + B_\kappa \le \mathrm{AURC}(S) + 2\mathfrak{E}^{\mathrm{sr}} + 2B_\kappa. \end{aligned}$$

Each individual base selector belongs to $\mathcal{S}$ by taking its weight to be one and the other weights to be zero. Taking the infimum over comparators satisfying the stated CDF bound gives the oracle bound with estimation error $2\mathfrak{E}^{\mathrm{sr}}$ from Eq. (B.1) and smoothing error $2B_\kappa = 6b_\kappa \log(e/b_\kappa) = O_{L_G}(\kappa \log(e/\kappa))$. □

B.2.2 Proof of Lemma 9 (smoothing error)

Definition 8 (Smoothed SoftRank-AURC plug-in objective)**.** *Write $\eta = \mathbb{I}\{y^z = -1\}$. For $\kappa > 0$, define the soft upper-tail probability and its self-inclusive empirical version by*

$$q_\kappa(\boldsymbol{x}; S) \doteq \mathbb{E}_{\boldsymbol{x}'}\sigma\left(\frac{S(\boldsymbol{x}') - S(\boldsymbol{x})}{\kappa}\right),$$
$$\widehat{q}_{\kappa,i}(S) \doteq \frac{1}{N}\sum_{j=1}^{N}\sigma\left(\frac{S(\boldsymbol{x}_j) - S(\boldsymbol{x}_i)}{\kappa}\right).$$

Equivalently, $q_\kappa = 1 - G_{S,\kappa}$ and $\widehat{q}_{\kappa,i} = 1 - \widehat{G}_\kappa(\boldsymbol{x}_i; S)$, with the CDF comparisons oriented as $\sigma((S(\boldsymbol{x}) - S(\boldsymbol{x}'))/\kappa)$. Define

$$\mathcal{R}_A(S) \doteq \mathbb{E}[-\eta \log q_\kappa(\boldsymbol{x}; S)],$$
$$\mathcal{L}_A(S) \doteq -\frac{1}{N}\sum_{i=1}^{N}\eta_i \log \widehat{q}_{\kappa,i}(S).$$

Here $\mathcal{L}_A$ is the objective in Eq. (3.10). Since the sum includes $j = i$, $\widehat{q}_{\kappa,i} \geq 1/(2N)$, so $\mathcal{L}_A$ is finite.

Lemma 9 (SoftRank-AURC smoothing approximation)**.** *Let S have a CDF F_S with Lipschitz constant at most L_G. For $0 < \kappa \leq 1$,*

$$\left|\mathrm{AURC}(S) - \mathcal{R}_A(S)\right| \leq B_\kappa \doteq 3b_\kappa \log(e/b_\kappa), \qquad b_\kappa = \min\{1, 2L_G\kappa\log 2\}.$$

Proof of Lemma 9. Write $q(\boldsymbol{x}) = 1 - F_S(S(\boldsymbol{x}))$. Since F_S is continuous, $q(\boldsymbol{x})$ is uniform on $(0, 1)$, and the population rank identity in Prop. 6 gives $\mathrm{AURC}(S) = \mathbb{E}[-\eta \log q(\boldsymbol{x})]$. The score distribution has a density bounded by L_G. For every score threshold s, integrating the absolute difference between the sigmoid and the upper-tail indicator yields

$$\left|\mathbb{E}\sigma\left(\frac{S(\boldsymbol{x}') - s}{\kappa}\right) - \Pr(S(\boldsymbol{x}') > s)\right| \leq L_G\int_{\mathbb{R}}\frac{du}{1 + e^{|u-s|/\kappa}} = 2L_G\kappa\log 2.$$

Also $q_\kappa(\boldsymbol{x}; S) \geq q(\boldsymbol{x})/2$, because $\sigma(v) \geq 1/2$ for $v \geq 0$. Consequently $|q_\kappa - q| \leq b_\kappa$. Put $b = b_\kappa$. On $q_\kappa \geq q$, $\log(q_\kappa/q) \leq \log(1 + b/q)$; on $q_\kappa < q$, $\log(q/q_\kappa) \leq \min\{\log 2, 2b/q\}$. Since $0 \leq \eta \leq 1$,

$$\begin{aligned}\left|\mathrm{AURC}(S) - \mathcal{R}_A(S)\right| &\leq \int_0^1 \log(1 + b/u)\, du + \int_0^1 \min\{\log 2, 2b/u\}\, du \\ &\leq (1 + b)\log(1 + b) - b\log b + b\log 2 + 2b\log(1/b) \\ &\leq 3b\log(e/b).\end{aligned}$$

The last inequality uses $(1 + b)\log(1 + b) \leq 2b$ for $0 < b \leq 1$. □

B.2.3 Proof of Lemma 10 (uniform deviation)

Lemma 10 (Uniform deviation for SoftRank-AURC)**.** *Fix $N \geq 2$, $\beta > 0$, $0 < \kappa \leq 1$, and $0 < \delta < 1$. For i.i.d. calibration observations and fixed base score maps, with probability at least $1 - \delta$,*

$$\sup_{S\in\mathcal{S}}\left|\mathcal{L}_A(S) - \mathcal{R}_A(S)\right| \leq \mathfrak{E}^{\mathrm{sr}}_{N,T,\delta}(\beta, \kappa),$$

where, for an absolute constant $C > 0$ and $\bar{\beta} = \min\{\beta, 1\}$,

$$\mathfrak{E}^{\mathrm{sr}}_{N,T,\delta}(\beta, \kappa) \doteq C\log(eN)\left[\sqrt{\frac{T\log\big(CN/(\bar{\beta}\kappa)\big) + \log(2/\delta)}{N}} + \frac{1}{N}\right]. \tag{B.1}$$

The same bound holds on $\overline{\mathcal{S}}$ by continuity. This statement does not require Assp. 1; in particular, score ties are allowed.

Proof of Lemma 10. We control the expected plug-in bias, concentration for a fixed score, and then the supremum over the score class. The argument retains the dependence between the inner and outer empirical averages. Constants denoted by C below are absolute and may increase between displays.

For a fixed score S, put $p(\boldsymbol{x}) = \Pr(S(\boldsymbol{x}') \geq S(\boldsymbol{x}))$. The upper-tail rank satisfies $\Pr(p(\boldsymbol{x}) \leq u) \leq u$, even with atoms. One way to see this is to introduce an independent $V \sim \mathrm{Unif}(0,1)$: the randomized tail rank $\Pr(S(\boldsymbol{x}') > S(\boldsymbol{x})) + V \Pr(S(\boldsymbol{x}') = S(\boldsymbol{x}))$ is uniform and is at most $p(\boldsymbol{x})$. As $q_\kappa \geq p/2$,

$$\Pr(q_\kappa(\boldsymbol{x}; S) \leq u) \leq 2u, \qquad 0 < u < 1. \tag{B.2}$$

In particular, $\mathcal{R}_A(S)$ is finite.

Condition on $(\boldsymbol{x}_i, \eta_i)$, and abbreviate $q = q_\kappa(\boldsymbol{x}_i; S)$. Then

$$\widehat{q}_{\kappa,i} = Z \doteq (1/2 + \sum\nolimits_{j=1}^{N-1} Y_j)/N, \qquad 0 \leq Y_j \leq 1, \quad \mathbb{E}Y_j = q,$$

where the Y_j's are conditionally independent. We claim that

$$|\mathbb{E}\log(Z/q)| \leq \begin{cases} C/(Nq), & Nq \geq 1, \\ \log(2N) - \log q, & Nq < 1. \end{cases} \tag{B.3}$$

For $Nq \geq 1$, Jensen's inequality gives $\mathbb{E}\log(Z/q) \leq \log(\mathbb{E}Z/q) \leq 1/(2Nq)$. To bound the other direction, write $n = N - 1$ and $\mu = nq$. If $\mu < 1 \leq Nq$, then $Nq < 2$, and $Z/q \geq 1/(2Nq) > 1/4$ suffices. If $\mu \geq 1$, coordinatewise convexity of $(1/2 + \sum_j y_j)^{-1}$ allows each Y_j to be replaced by a Bernoulli variable of mean q for an upper bound on its expectation. For $B \sim \mathrm{Bin}(n, q)$,

$$\begin{aligned} \mathbb{E}\frac{1}{1/2 + \sum_j Y_j} \leq \mathbb{E}\frac{1}{1/2 + B} &= \int_0^1 t^{-1/2}(1 - q + qt)^n \, dt \\ &\leq \int_0^1 (1-v)^{-1/2} e^{-\mu v} \, dv \leq \frac{1}{\mu} + \frac{C}{\mu^2}. \end{aligned}$$

For the final estimate, split the integral at $v = 1/2$, use $(1-v)^{-1/2} \leq 1 + 2v$ on the first part, and bound the second part by $\sqrt{2}e^{-\mu/2} \leq C/\mu^2$. The inequality $\log u \geq 1 - 1/u$, together with $n \geq N/2$, now gives $\mathbb{E}\log(Z/q) \geq 1 - Nq(1/\mu + C/\mu^2) \geq -C/(Nq)$. For $Nq < 1$, use $1/(2N) \leq Z \leq 1$ to obtain the second line of Eq. (B.3).

Integrating Eq. (B.3) with Eq. (B.2) gives

$$\begin{aligned} \mathbb{E}\big[q^{-1}\mathbb{I}\{q \geq 1/N\}\big] &\leq 1 + 2\log N, \\ \mathbb{E}[(\log(2N) - \log q)\mathbb{I}\{q < 1/N\}] &\leq C\frac{\log(eN)}{N}. \end{aligned}$$

Since η_i is fixed under the conditioning and belongs to $[0,1]$, exchangeability yields the distribution-free bias bound

$$\sup\nolimits_{S \in \mathcal{S}} |\mathbb{E}\mathcal{L}_A(S) - \mathcal{R}_A(S)| \leq C\log(eN)/N. \tag{B.4}$$

For a fixed S, replace one observation in the calibration sample. Its own summand in $\mathcal{L}_A$ changes by at most $\log(2N)/N$. Order the other $N-1$ common observations by decreasing score, breaking ties arbitrarily, and let $r = 1, \ldots, N-1$ denote this order. The observation of rank r has at least r common comparisons with sigmoid value at least $1/2$, including its self-comparison. Thus both of its empirical tail probabilities are at least $r/(2N)$, whereas replacing one reference point changes that probability by at most $1/N$. Its log weight therefore changes by at most $2/r$. The entire loss changes by at most

$$d_N/N, \qquad d_N = \log(2N) + 2\sum\nolimits_{r=1}^{N-1} \frac{1}{r} \leq C\log(eN).$$

McDiarmid's inequality consequently gives

$$\Pr(|\mathcal{L}_A(S) - \mathbb{E}\mathcal{L}_A(S)| > t) \leq 2\exp\big(-2Nt^2/d_N^2\big). \tag{B.5}$$

We now extend the concentration bound to all scores in $\mathcal{S}$. Since $|(\log\sigma)'| \le 1$, a perturbation $\|S - S'\|_\infty \le r$ changes each log pairwise comparison by at most $2r/\kappa$. The same multiplicative bounds hold after averaging, so

$$|\mathcal{L}_A(S) - \mathcal{L}_A(S')| \le 2r/\kappa, \qquad |\mathcal{R}_A(S) - \mathcal{R}_A(S')| \le 2r/\kappa.$$

Take an $r = \kappa/(8N)$ net in $\|\cdot\|_\infty$. By Lem. 11, its cardinality M satisfies $\log M \le CT\log\big(CN/(\bar\beta\kappa)\big)$. A union bound in Eq. (B.5), the bias estimate Eq. (B.4), and extension from the net give, with probability at least $1-\delta$,

$$\sup_{S\in\mathcal{S}} |\mathcal{L}_A(S) - \mathcal{R}_A(S)| \le d_N\sqrt{\frac{\log(2M/\delta)}{2N}} + C\frac{\log(eN)}{N} + \frac{4r}{\kappa} \le \mathfrak{E}^{\mathrm{sr}}_{N,T,\delta}(\beta,\kappa).$$

No CDF regularity was used. □

B.2.4 Proof of Lemma 11 (covering number)

Lemma 11 (Entropy of the selector class)**.** *There exist absolute constants $C_0, C_1 > 0$ such that for every $\beta > 0$ and every $0 < \varepsilon \le 1$,*

$$\log\mathcal{N}\big(\varepsilon, \mathcal{S}, \|\cdot\|_\infty\big) \le C_0 T \log\Big(\frac{C_1}{\bar\beta\,\varepsilon}\Big), \qquad \bar\beta := \min\{\beta, 1\}.$$

Proof of Lemma 11. Write

$$S(\boldsymbol{x}) = \sum\nolimits_{t=1}^{T} \omega_t g_{t,\gamma_t}(\boldsymbol{x}), \qquad g_{t,\gamma}(\boldsymbol{x}) := \tanh\Big(\frac{s_t(\boldsymbol{x}) - \gamma}{\beta}\Big).$$

Fix $0 < \varepsilon \le 1$ and write $\bar\beta = \min\{\beta, 1\}$. Since $|\partial_\gamma g_{t,\gamma}(\boldsymbol{x})| \le 1/\bar\beta$ uniformly in $(\boldsymbol{x}, \gamma)$, for each $t \in [T]$ there exists a finite set $\Gamma_{t,\varepsilon} \subset [0,1]$ with

$$|\Gamma_{t,\varepsilon}| \le 1 + \frac{2}{\bar\beta\varepsilon}$$

such that for every $\gamma \in [0,1]$ one can find $\widetilde\gamma \in \Gamma_{t,\varepsilon}$ satisfying

$$\|g_{t,\gamma} - g_{t,\widetilde\gamma}\|_\infty \le \varepsilon/2.$$

Let $\Omega_\varepsilon \subset \Delta^T$ be an ℓ_1-net of the Δ^T at radius $\varepsilon/2$, chosen so that

$$|\Omega_\varepsilon| \le (C/\varepsilon)^T$$

for an absolute constant $C > 0$; this is the standard volumetric bound for $\Delta^T \subset \mathbb{R}^T$.

Consider the finite class

$$\widetilde{\mathcal{S}}_\varepsilon := \Big\{ \sum\nolimits_{t=1}^{T} \widetilde\omega_t g_{t,\widetilde\gamma_t}(\boldsymbol{x}) : \widetilde{\boldsymbol{\omega}} \in \Omega_\varepsilon,\ \widetilde\gamma_t \in \Gamma_{t,\varepsilon} \text{ for all } t \Big\}.$$

Take any $S \in \mathcal{S}$, represented by $(\boldsymbol{\omega}, \boldsymbol{\gamma})$. Choose $\widetilde{\boldsymbol{\omega}} \in \Omega_\varepsilon$ with $\|\boldsymbol{\omega} - \widetilde{\boldsymbol{\omega}}\|_1 \le \varepsilon/2$, and for each t choose $\widetilde\gamma_t \in \Gamma_{t,\varepsilon}$ with $\|g_{t,\gamma_t} - g_{t,\widetilde\gamma_t}\|_\infty \le \varepsilon/2$. Then, since $\sum_t \omega_t \le 1$ and $|g_{t,\widetilde\gamma_t}| \le 1$,

$$\begin{aligned}
\Big\| S - \sum\nolimits_{t=1}^{T} \widetilde\omega_t g_{t,\widetilde\gamma_t} \Big\|_\infty &\le \Big\| \sum\nolimits_{t=1}^{T} \omega_t \big(g_{t,\gamma_t} - g_{t,\widetilde\gamma_t}\big) \Big\|_\infty + \Big\| \sum\nolimits_{t=1}^{T} (\omega_t - \widetilde\omega_t) g_{t,\widetilde\gamma_t} \Big\|_\infty \\
&\le \sum\nolimits_{t=1}^{T} \omega_t \|g_{t,\gamma_t} - g_{t,\widetilde\gamma_t}\|_\infty + \sum\nolimits_{t=1}^{T} |\omega_t - \widetilde\omega_t| \\
&\le \varepsilon/2 + \varepsilon/2 = \varepsilon.
\end{aligned}$$

Thus $\widetilde{\mathcal{S}}_\varepsilon$ is an ε-net of $\mathcal{S}$ in $\|\cdot\|_\infty$. Its cardinality satisfies

$$\mathcal{N}(\varepsilon, \mathcal{S}, \|\cdot\|_\infty) \le |\widetilde{\mathcal{S}}_\varepsilon| \le (C/\varepsilon)^T \big(1 + 2/\big(\bar\beta\varepsilon\big)\big)^T.$$

Taking logarithms and using $0 < \bar\beta, \varepsilon \le 1$ yields

$$\log\mathcal{N}\big(\varepsilon, \mathcal{S}, \|\cdot\|_\infty\big) \le C_0 T \log\big[C_1/\big(\bar\beta\,\varepsilon\big)\big]$$

for suitable absolute constants $C_0, C_1 > 0$. □

C Additional Experimental Results

Post-Hoc Confidence Scores The ten post-hoc confidence scores and their references are summarized in Tab. 1 in the main text. For KNN, ViM, and SIRC, which require additional hyperparameters, we tune them on the calibration split. As in [4], we set the KNN neighborhood size to $k = 2$ unless otherwise specified.

Table 5: Dataset, model, and LLM benchmark statistics in transposed format. Ex. denotes the number of evaluation-pool examples for image/NLP datasets and the number of filtered benchmark examples for LLM tasks; K is the number of classes or answer options. Image/NLP column abbreviations follow Tab. 2. Model abbreviations are R32 for resnet32, R50 for resnet50, LTC for LONGTAILCXR, and M/Q for the Mistral-7B-v0.1 and Qwen2.5-7B LLM backbones. HF denotes Hugging Face, including the LLM sources mistralai/Mistral-7B-v0.1 and Qwen/Qwen2.5-7B, and Code denotes official source code.

Field	ADE	AMZ	CAL	C10	C1C	C100	EYE	IMN	INO	WID	WOD	OIO	NIH	PCM	SST	MMLU	MMLU Pro	Hella Swag	ARC Easy
Ex.	4, 704	100, 050	1, 395	10, 000	190, 000	20, 000	53, 576	50, 000	52, 000	8, 154	42, 791	67, 632	20, 279	32, 768	1, 821	14, 042	9, 981	10, 042	2, 365
K	5	5	101	10	10	100	5	1000	1000	182	182	1000	20	2	2	4	10	4	4
Model	spanbert	LISA	FLYP	R32	R32	R32	–	R50	R50	FLYP	FLYP	R50	LTC	FLYP	FLYP	M/Q	M/Q	M/Q	M/Q
Source	HF	Code	Code	timm	timm	timm	Code	timm	timm	Code	Code	timm	Code	Code	Code	HF	HF	HF	HF

Datasets and Models For SpanBERT in Tab. 5, we use the SpanBERT model [47] fine-tuned on the ADE corpus (Huggingface model ID: abhibisht89/spanbert-large-cased-finetuned-ade_corpus_v2).

For each LLM example, let $\boldsymbol{z}_i \in \mathbb{R}^K$ denote the logits of the K answer-option tokens at the final prompt position. Following the observation in [4] that softmax-response confidence scores are sensitive to the raw-logit scale, we compute all LLM confidence scores from the sample-wise scaled logits

$$\tilde{\boldsymbol{z}}_i = \boldsymbol{z}_i K / \sum_{j=1}^{K} |z_{ij}|.$$

This rescaling prevents very large LLM logits from making a single option dominate the softmax responses, while the positive per-sample scale preserves the candidate-token argmax.

Detailed Calibration-Size Results **Detailed calibration-size results for the image/NLP datasets** are reported from Tab. 9 to Tab. 21. All entries are AURC values multiplied by 100 and averaged over 5 random seeds; lower is better. The *Cali* column gives the calibration size $N \in \{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$, and dataset columns use the abbreviations from Tab. 2. Rows include AURC-MLP and the ensemble variants whose learned parameters can change with calibration size. **Green** marks best or tied-best values for each dataset and calibration size, and **Blue** marks second-best or tied second-best values at displayed precision. **Full LLM calibration-size** sweeps are reported in Tabs. 22 to 25. All entries are AURC values multiplied by 100 and reported as mean (standard deviation) over 5 random seeds. Columns correspond to calibration size $N \in \{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$. Fixed raw-logit, softmax-response, energy, and margin baselines do not depend on calibration size and are repeated across columns for direct comparison. For smooth ensemble variants, the inner β is selected on the validation set; for direct-AURC variants, κ is independently grid-searched over the same range as β. **Green** marks best or tied-best values within each column, and **Blue** marks second-best or tied second-best values at displayed precision.

Learned ensemble weights. To understand how the base scorers are combined, we examine the weights learned by AURC-SCR and AURC-SLR. Tab. 6 reports the normalized weight assigned to each base scorer on each of the 15 image and NLP datasets, together with the share held by the three largest weights (top-3). We report weights learned with $N = 1000$ calibration examples under a setting that differs from the main protocol in Sec. 4. We highlight two observations.

We see a trend of assigning larger weights to better-performing selectors on multiple datasets. For example, on ADE, KNN is the strongest individual scorer and receives the largest weight under AURC-SCR. The weight of the best individual scorer never falls below 1%. However, the weights are

not determined by performance alone. For example, on Caltech101 under AURC-SLR, Energy is 2.4% worse than SR_{ent}. Yet Energy receives the largest weight in that combination, while SR_{ent} receives only 5.6%. The final ensemble performance is still better than any individual scorer, suggesting the ensemble is exploiting complementary information from the base scorers.

In many datasets, the weights are concentrated on a relatively small subset of scorers. KNN, SIRC, $RL_{\text{conf-M}}$, and $RL_{\text{geo-M}}$ are frequently among the scorers receiving the largest weights. But the selectors with the large weights may vary across datasets. There is no single base selector that is assigned high weights for all datasets. Similarly, no single scorer is consistently inactive across all datasets and objectives, suggesting that the usefulness of an individual scorer depends on the dataset and optimization criterion.

Table 6: Normalized ensemble weights (%) learned by AURC-SCR and AURC-SLR at calibration size $N = 1000$. Top-3 is the share held by the three largest weights. The row #off counts the datasets (out of 15) on which a scorer receives less than 1% of the weight.

Variant	Data	RL_{max}	SR_{max}	$\text{SR}_{\text{doctor}}$	SR_{ent}	ViM	SIRC	KNN	Energy	$\text{RL}_{\text{conf-M}}$	$\text{RL}_{\text{geo-M}}$	Top-3
AURC-SCR	ADE	0.2	0.1	0.2	0.4	0.0	45.9	52.9	0.0	0.1	0.1	99.3
	AMZ	5.7	13.1	12.3	35.8	7.7	19.5	2.6	2.0	1.3	0.0	86.4
	CAL	16.1	2.1	23.6	35.4	1.5	3.1	0.1	13.8	2.2	2.1	77.4
	C10	0.6	4.8	0.8	0.3	11.7	0.3	34.9	0.4	33.5	12.8	88.2
	C1C	3.1	0.3	0.0	0.1	0.2	0.1	56.4	3.4	31.7	4.7	95.7
	C100	0.0	0.4	0.3	0.7	0.2	0.1	85.0	0.0	6.5	6.7	98.7
	EYE	0.0	21.3	21.5	3.3	12.4	15.7	7.6	0.0	12.3	5.8	75.7
	IMN	7.2	6.0	3.6	2.6	0.2	0.1	2.9	3.2	36.7	37.4	90.3
	INO	13.1	4.6	3.3	0.5	0.0	0.1	1.8	13.6	32.0	31.0	81.6
	WID	0.2	16.0	11.5	6.8	0.0	8.9	19.7	0.2	18.4	18.4	65.7
	WOD	0.1	16.0	12.7	11.2	3.3	1.6	33.2	0.1	11.0	11.0	72.9
	OIO	24.9	6.3	9.2	4.3	1.0	1.1	1.6	16.6	14.8	20.2	64.5
	NIH	4.5	0.0	1.1	4.6	0.4	2.1	49.6	32.4	1.0	4.4	93.5
	PCM	0.0	6.1	6.0	4.7	21.9	9.4	18.6	0.0	16.6	16.6	74.5
	SST	11.9	2.0	2.1	2.6	15.3	13.4	0.7	6.7	22.6	22.6	69.3
	#off	7	4	4	5	8	5	2	7	2	2	–
AURC-SLR	ADE	7.6	3.3	4.1	1.7	0.0	36.6	34.6	6.4	3.2	2.4	80.8
	AMZ	4.9	13.7	14.8	36.4	3.9	15.2	4.7	2.6	3.3	0.6	75.4
	CAL	12.3	12.1	11.4	5.6	2.2	10.5	0.4	16.4	14.6	14.6	58.7
	C10	2.9	8.7	0.4	1.4	20.5	1.8	23.0	3.1	20.3	17.8	77.5
	C1C	13.0	4.4	4.1	0.3	0.0	2.7	16.9	21.9	26.4	10.3	82.9
	C100	1.4	3.0	3.0	3.6	3.2	0.1	52.0	6.6	13.2	13.9	87.1
	EYE	0.0	20.7	19.5	0.7	14.2	14.1	7.0	0.4	15.9	7.3	80.2
	IMN	6.2	5.3	3.5	2.9	0.1	0.3	3.1	4.1	36.6	38.0	88.5
	INO	12.8	13.0	10.1	1.7	2.2	2.3	3.3	13.3	19.9	21.4	58.6
	WID	0.2	29.3	6.3	4.0	0.0	7.7	20.2	0.0	16.1	16.1	72.8
	WOD	1.0	14.1	13.7	12.8	1.5	12.5	17.1	1.9	13.0	12.3	45.4
	OIO	13.0	12.3	12.7	12.0	5.4	9.5	1.2	12.7	10.5	10.7	39.4
	NIH	2.5	0.0	1.4	5.0	3.2	0.0	57.7	25.6	0.0	4.7	93.6
	PCM	9.5	4.6	4.8	5.2	42.5	9.9	3.0	3.4	8.6	8.6	68.4
	SST	20.6	6.8	7.5	3.2	26.8	9.4	9.1	6.6	5.0	5.0	73.3
	#off	3	1	1	2	4	3	1	2	1	1	–

Ablation Study of the Number of Base Selectors The full cross-dataset ablation results for varying the number of base selectors are reported in Tab. 26.

D Compute Resources

All experiments were run on an internal university SLURM cluster. The compute used by each stage of the pipeline is summarized in Tab. 27. The proposed ensemble training and risk–coverage evaluation are CPU-only once fixed-model logits, features, and base confidence scores have been cached.

For the image/NLP datasets in Sec. 4.1, collecting fixed-model logits, features, and base confidence scores uses the same single-GPU collection setup and requires at most 45 A100-GPU-hours across the 15 dataset/model pairs. **For the LLM benchmarks** in Sec. 4.2, collecting option-token logits for 4 benchmarks and 2 open-weight 7B models requires at most 24 A100-GPU-hours. Thus, the final reported feature/logit collection stage uses at most 69 A100-GPU-hours in total. The remaining experiments train post-hoc ensemble selectors on cached scores. Running the full hyperparameter

Table 7: Detailed calibration-size sweep on ADE. Entries are mean (standard deviation) over 5 repeats. **Blue**-background rows are our methods. **Green** numbers indicate the best mean score at each calibration size. **Blue** numbers indicate the second-best mean.

	ADE									
Method	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
RL_{max}	87.53 (0.75)	87.53 (0.75)	87.53 (0.75)	87.53 (0.75)	87.53 (0.75)	87.53 (0.75)	87.53 (0.75)	87.53 (0.75)	87.53 (0.75)	87.53 (0.75)
SR_{max}	87.06 (0.74)	87.06 (0.74)	87.06 (0.74)	87.06 (0.74)	87.06 (0.74)	87.06 (0.74)	87.06 (0.74)	87.06 (0.74)	87.06 (0.74)	87.06 (0.74)
SR_{doctor}	85.39 (0.81)	85.39 (0.81)	85.39 (0.81)	85.39 (0.81)	85.39 (0.81)	85.39 (0.81)	85.39 (0.81)	85.39 (0.81)	85.39 (0.81)	85.39 (0.81)
SR_{ent}	85.25 (0.81)	85.25 (0.81)	85.25 (0.81)	85.25 (0.81)	85.25 (0.81)	85.25 (0.81)	85.25 (0.81)	85.25 (0.81)	85.25 (0.81)	85.25 (0.81)
$Energy$	83.13 (0.68)	83.13 (0.68)	83.13 (0.68)	83.13 (0.68)	83.13 (0.68)	83.13 (0.68)	83.13 (0.68)	83.13 (0.68)	83.13 (0.68)	83.13 (0.68)
$RL_{\text{conf-M}}$	87.64 (0.82)	87.64 (0.82)	87.64 (0.82)	87.64 (0.82)	87.64 (0.82)	87.64 (0.82)	87.64 (0.82)	87.64 (0.82)	87.64 (0.82)	87.64 (0.82)
$RL_{\text{geo-M}}$	88.20 (0.80)	88.20 (0.80)	88.20 (0.80)	88.20 (0.80)	88.20 (0.80)	88.20 (0.80)	88.20 (0.80)	88.20 (0.80)	88.20 (0.80)	88.20 (0.80)
ViM	71.68 (2.03)	71.28 (2.43)	72.83 (1.10)	73.00 (0.92)	72.99 (1.15)	73.36 (1.20)	73.63 (1.58)	73.83 (1.65)	73.76 (1.94)	73.62 (2.12)
$SIRC$	56.16 (1.55)	56.33 (1.65)	55.79 (1.81)	55.97 (1.51)	55.99 (1.60)	55.87 (1.35)	56.10 (1.51)	56.13 (1.83)	56.25 (1.39)	56.53 (1.17)
KNN	**55.88** (1.52)	55.56 (1.56)	55.47 (1.21)	55.40 (1.37)	55.27 (1.39)	55.12 (1.12)	55.19 (1.07)	55.25 (1.05)	55.09 (0.94)	55.18 (1.00)
AURC-MLP	57.04 (3.62)	55.34 (3.66)	55.58 (2.74)	55.29 (2.83)	55.88 (2.74)	55.81 (2.76)	55.31 (3.09)	55.53 (2.46)	55.45 (2.71)	55.82 (1.47)
SELE-SCR	60.22 (6.18)	55.92 (1.51)	55.48 (1.36)	55.45 (1.58)	55.44 (1.59)	55.53 (1.45)	55.49 (1.40)	55.47 (1.44)	55.59 (1.37)	55.50 (1.34)
SELE-SLR	**54.86** (1.60)	**53.48** (1.15)	**53.58** (1.47)	**53.27** (1.50)	**53.26** (1.64)	**52.79** (1.39)	**53.54** (1.21)	**53.24** (1.56)	**53.39** (1.40)	**53.16** (1.23)
AURC-SCR	59.51 (4.90)	55.78 (1.31)	55.60 (1.69)	55.56 (1.43)	55.57 (1.58)	55.00 (1.57)	55.19 (1.50)	55.40 (1.44)	55.61 (1.54)	55.25 (1.67)
AURC-SLR	55.93 (2.67)	**53.63** (0.67)	**53.35** (1.87)	**52.99** (1.60)	**53.18** (1.75)	**52.96** (1.20)	**53.66** (0.83)	**52.76** (1.18)	**53.26** (1.90)	**53.21** (1.81)

Table 8: Detailed calibration-size sweep on AMAZON. Entries are mean (standard deviation) over 5 repeats. **Blue**-background rows are our methods. **Green** numbers indicate the best mean score at each calibration size. **Blue** numbers indicate the second-best mean.

	AMAZON									
Method	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
RL_{max}	12.55 (0.21)	12.55 (0.21)	12.55 (0.21)	12.55 (0.21)	12.55 (0.21)	12.55 (0.21)	12.55 (0.21)	12.55 (0.21)	12.55 (0.21)	12.55 (0.21)
SR_{max}	12.32 (0.24)	12.32 (0.24)	12.32 (0.24)	12.32 (0.24)	12.32 (0.24)	12.32 (0.24)	12.32 (0.24)	12.32 (0.24)	12.32 (0.24)	12.32 (0.24)
SR_{doctor}	12.27 (0.24)	12.27 (0.24)	12.27 (0.24)	12.27 (0.24)	12.27 (0.24)	12.27 (0.24)	12.27 (0.24)	12.27 (0.24)	12.27 (0.24)	12.27 (0.24)
SR_{ent}	**12.13** (0.23)	**12.13** (0.23)	**12.13** (0.23)	**12.13** (0.23)	**12.13** (0.23)	**12.13** (0.23)	**12.13** (0.23)	**12.13** (0.23)	**12.13** (0.23)	**12.13** (0.23)
$Energy$	12.86 (0.21)	12.86 (0.21)	12.86 (0.21)	12.86 (0.21)	12.86 (0.21)	12.86 (0.21)	12.86 (0.21)	12.86 (0.21)	12.86 (0.21)	12.86 (0.21)
$RL_{\text{conf-M}}$	12.56 (0.25)	12.56 (0.25)	12.56 (0.25)	12.56 (0.25)	12.56 (0.25)	12.56 (0.25)	12.56 (0.25)	12.56 (0.25)	12.56 (0.25)	12.56 (0.25)
$RL_{\text{geo-M}}$	12.87 (0.26)	12.87 (0.26)	12.87 (0.26)	12.87 (0.26)	12.87 (0.26)	12.87 (0.26)	12.87 (0.26)	12.87 (0.26)	12.87 (0.26)	12.87 (0.26)
ViM	14.58 (0.31)	14.51 (0.26)	14.49 (0.27)	14.47 (0.26)	14.43 (0.25)	14.41 (0.23)	14.40 (0.22)	14.40 (0.22)	14.40 (0.22)	14.40 (0.22)
$SIRC$	**12.23** (0.24)	12.23 (0.24)	12.23 (0.24)	12.23 (0.24)	12.23 (0.24)	12.23 (0.24)	12.23 (0.24)	12.23 (0.24)	12.23 (0.24)	12.23 (0.24)
KNN	18.13 (0.59)	18.41 (0.47)	18.34 (0.34)	18.42 (0.32)	18.37 (0.29)	18.29 (0.41)	18.38 (0.38)	18.38 (0.46)	18.42 (0.45)	18.38 (0.42)
AURC-MLP	19.41 (3.16)	18.78 (2.21)	17.88 (1.34)	17.45 (1.08)	17.40 (1.29)	17.28 (0.83)	16.93 (0.50)	16.83 (0.33)	17.13 (0.58)	16.92 (0.35)
SELE-SCR	12.35 (0.18)	12.28 (0.27)	12.25 (0.22)	12.20 (0.25)	12.19 (0.21)	12.17 (0.22)	12.17 (0.23)	12.17 (0.23)	12.15 (0.24)	12.15 (0.24)
SELE-SLR	12.32 (0.18)	12.29 (0.28)	12.24 (0.21)	**12.19** (0.23)	**12.17** (0.23)	**12.15** (0.24)	12.16 (0.22)	**12.14** (0.23)	**12.12** (0.23)	**12.13** (0.25)
AURC-SCR	12.34 (0.18)	12.40 (0.40)	12.26 (0.32)	12.22 (0.31)	12.22 (0.32)	**12.15** (0.25)	12.15 (0.24)	12.15 (0.26)	**12.12** (0.25)	12.14 (0.25)
AURC-SLR	12.33 (0.17)	12.35 (0.24)	12.21 (0.25)	**12.19** (0.26)	12.18 (0.25)	**12.13** (0.24)	**12.14** (0.23)	**12.13** (0.25)	**12.12** (0.24)	**12.12** (0.26)

Table 9: Detailed calibration-size sweep on CALTECH. Entries are mean (standard deviation) over 5 repeats. **Blue**-background rows are our methods. **Green** numbers indicate the best mean score at each calibration size. **Blue** numbers indicate the second-best mean.

	CALTECH									
Method	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
RL_{max}	2.21 (0.52)	2.21 (0.52)	2.21 (0.52)	2.21 (0.52)	2.21 (0.52)	2.21 (0.52)	2.21 (0.52)	2.21 (0.52)	2.21 (0.52)	2.21 (0.52)
SR_{max}	1.01 (0.17)	1.01 (0.17)	1.01 (0.17)	1.01 (0.17)	1.01 (0.17)	1.01 (0.17)	1.01 (0.17)	1.01 (0.17)	1.01 (0.17)	1.01 (0.17)
SR_{doctor}	1.01 (0.17)	1.01 (0.17)	1.01 (0.17)	1.01 (0.17)	1.01 (0.17)	1.01 (0.17)	1.01 (0.17)	1.01 (0.17)	1.01 (0.17)	1.01 (0.17)
SR_{ent}	0.98 (0.16)	0.98 (0.16)	0.98 (0.16)	0.98 (0.16)	0.98 (0.16)	0.98 (0.16)	0.98 (0.16)	0.98 (0.16)	0.98 (0.16)	0.98 (0.16)
$Energy$	2.37 (0.56)	2.37 (0.56)	2.37 (0.56)	2.37 (0.56)	2.37 (0.56)	2.37 (0.56)	2.37 (0.56)	2.37 (0.56)	2.37 (0.56)	2.37 (0.56)
$RL_{\text{conf-M}}$	1.27 (0.23)	1.27 (0.23)	1.27 (0.23)	1.27 (0.23)	1.27 (0.23)	1.27 (0.23)	1.27 (0.23)	1.27 (0.23)	1.27 (0.23)	1.27 (0.23)
$RL_{\text{geo-M}}$	1.27 (0.23)	1.27 (0.23)	1.27 (0.23)	1.27 (0.23)	1.27 (0.23)	1.27 (0.23)	1.27 (0.23)	1.27 (0.23)	1.27 (0.23)	1.27 (0.23)
ViM	6.85 (0.91)	5.71 (0.43)	5.25 (0.44)	4.94 (0.37)	4.51 (0.19)	4.20 (0.30)	4.06 (0.37)	4.06 (0.37)	4.06 (0.37)	4.06 (0.37)
$SIRC$	1.03 (0.18)	1.03 (0.18)	1.03 (0.17)	1.03 (0.17)	1.04 (0.17)	1.04 (0.17)	1.04 (0.17)	1.04 (0.17)	1.04 (0.17)	1.04 (0.17)
KNN	8.04 (1.20)	6.92 (0.59)	6.36 (0.59)	6.24 (0.85)	5.84 (0.77)	5.59 (0.75)	5.46 (0.73)	5.46 (0.73)	5.46 (0.73)	5.46 (0.73)
AURC-MLP	3.98 (1.15)	3.40 (0.58)	3.88 (1.35)	3.20 (1.51)	3.48 (1.03)	3.45 (0.58)	3.41 (0.82)	3.41 (0.82)	3.41 (0.82)	3.41 (0.82)
SELE-SCR	**0.74** (0.11)	0.78 (0.17)	0.76 (0.12)	0.74 (0.11)	0.76 (0.15)	0.77 (0.16)	0.75 (0.13)	0.75 (0.13)	0.75 (0.13)	0.75 (0.13)
SELE-SLR	**0.74** (0.11)	**0.74** (0.16)	**0.74** (0.12)	**0.72** (0.12)	**0.73** (0.14)	**0.74** (0.13)	**0.72** (0.13)	**0.72** (0.13)	**0.72** (0.13)	**0.72** (0.13)
AURC-SCR	0.75 (0.12)	0.78 (0.17)	0.78 (0.11)	0.74 (0.11)	0.76 (0.16)	0.77 (0.16)	0.73 (0.11)	0.73 (0.11)	0.73 (0.11)	0.73 (0.11)
AURC-SLR	**0.69** (0.08)	**0.70** (0.11)	**0.67** (0.13)	**0.66** (0.10)	**0.68** (0.12)	**0.67** (0.11)	**0.68** (0.13)	**0.68** (0.13)	**0.68** (0.13)	**0.68** (0.13)

Table 10: Detailed calibration-size sweep on CIFAR10. Entries are mean (standard deviation) over 5 repeats. **Blue**-background rows are our methods. **Green** numbers indicate the best mean score at each calibration size. **Blue** numbers indicate the second-best mean.

	CIFAR10									
Method	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
RL_{max}	1.64 (0.16)	1.64 (0.16)	1.64 (0.16)	1.64 (0.16)	1.64 (0.16)	1.64 (0.16)	1.64 (0.16)	1.64 (0.16)	1.64 (0.16)	1.64 (0.16)
SR_{max}	1.45 (0.12)	1.45 (0.12)	1.45 (0.12)	1.45 (0.12)	1.45 (0.12)	1.45 (0.12)	1.45 (0.12)	1.45 (0.12)	1.45 (0.12)	1.45 (0.12)
SR_{doctor}	1.45 (0.12)	1.45 (0.12)	1.45 (0.12)	1.45 (0.12)	1.45 (0.12)	1.45 (0.12)	1.45 (0.12)	1.45 (0.12)	1.45 (0.12)	1.45 (0.12)
SR_{ent}	1.47 (0.13)	1.47 (0.13)	1.47 (0.13)	1.47 (0.13)	1.47 (0.13)	1.47 (0.13)	1.47 (0.13)	1.47 (0.13)	1.47 (0.13)	1.47 (0.13)
$Energy$	1.64 (0.16)	1.64 (0.16)	1.64 (0.16)	1.64 (0.16)	1.64 (0.16)	1.64 (0.16)	1.64 (0.16)	1.64 (0.16)	1.64 (0.16)	1.64 (0.16)
$RL_{\text{conf-M}}$	1.05 (0.08)	1.05 (0.08)	1.05 (0.08)	1.05 (0.08)	1.05 (0.08)	1.05 (0.08)	1.05 (0.08)	1.05 (0.08)	1.05 (0.08)	1.05 (0.08)
$RL_{\text{geo-M}}$	1.10 (0.05)	1.10 (0.05)	1.10 (0.05)	1.10 (0.05)	1.10 (0.05)	1.10 (0.05)	1.10 (0.05)	1.10 (0.05)	1.10 (0.05)	1.10 (0.05)
ViM	1.25 (0.19)	1.13 (0.10)	1.11 (0.07)	1.09 (0.06)	1.10 (0.07)	1.07 (0.07)	1.07 (0.07)	1.05 (0.07)	1.06 (0.07)	1.06 (0.07)
$SIRC$	1.43 (0.13)	1.42 (0.12)	1.42 (0.12)	1.42 (0.11)	1.42 (0.12)	1.42 (0.12)	1.42 (0.12)	1.42 (0.12)	1.42 (0.12)	1.42 (0.12)
KNN	1.14 (0.32)	**0.98** (0.08)	**0.97** (0.05)	**0.98** (0.06)	0.98 (0.06)	0.99 (0.06)	1.00 (0.07)	1.01 (0.07)	1.01 (0.07)	1.00 (0.07)
AURC-MLP	4.07 (0.98)	3.19 (0.60)	3.32 (0.81)	2.85 (0.70)	2.66 (0.93)	2.04 (0.31)	3.42 (2.60)	2.16 (0.38)	2.19 (0.44)	2.27 (0.49)
SELE-SCR	**1.02** (0.08)	**0.98** (0.10)	**0.98** (0.08)	**0.97** (0.09)	**0.97** (0.10)	**0.97** (0.10)	**0.99** (0.11)	**0.99** (0.11)	**0.99** (0.12)	**0.98** (0.12)
SELE-SLR	**1.01** (0.07)	**0.99** (0.10)	**0.98** (0.08)	**0.98** (0.10)	**0.96** (0.09)	**0.97** (0.09)	**0.99** (0.10)	**0.99** (0.08)	**0.98** (0.10)	**0.98** (0.11)
AURC-SCR	1.05 (0.06)	1.02 (0.07)	**0.98** (0.05)	**0.98** (0.06)	**0.97** (0.07)	**0.98** (0.07)	**0.98** (0.08)	**0.99** (0.08)	**0.98** (0.08)	**0.97** (0.07)
AURC-SLR	1.05 (0.07)	1.01 (0.08)	**0.98** (0.06)	**0.98** (0.06)	**0.97** (0.07)	**0.98** (0.08)	**0.98** (0.08)	**0.98** (0.08)	**0.98** (0.08)	**0.97** (0.07)

Table 11: Detailed calibration-size sweep on CIFAR10-C. Entries are mean (standard deviation) over 5 repeats. **Blue**-background rows are our methods. **Green** numbers indicate the best mean score at each calibration size. **Blue** numbers indicate the second-best mean.

	CIFAR10-C									
Method	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
RL_{max}	7.49 (0.10)	7.49 (0.10)	7.49 (0.10)	7.49 (0.10)	7.49 (0.10)	7.49 (0.10)	7.49 (0.10)	7.49 (0.10)	7.49 (0.10)	7.49 (0.10)
SR_{max}	7.36 (0.09)	7.36 (0.09)	7.36 (0.09)	7.36 (0.09)	7.36 (0.09)	7.36 (0.09)	7.36 (0.09)	7.36 (0.09)	7.36 (0.09)	7.36 (0.09)
SR_{doctor}	7.36 (0.09)	7.36 (0.09)	7.36 (0.09)	7.36 (0.09)	7.36 (0.09)	7.36 (0.09)	7.36 (0.09)	7.36 (0.09)	7.36 (0.09)	7.36 (0.09)
SR_{ent}	7.37 (0.09)	7.37 (0.09)	7.37 (0.09)	7.37 (0.09)	7.37 (0.09)	7.37 (0.09)	7.37 (0.09)	7.37 (0.09)	7.37 (0.09)	7.37 (0.09)
$Energy$	7.48 (0.10)	7.48 (0.10)	7.48 (0.10)	7.48 (0.10)	7.48 (0.10)	7.48 (0.10)	7.48 (0.10)	7.48 (0.10)	7.48 (0.10)	7.48 (0.10)
$RL_{\text{conf-M}}$	6.80 (0.08)	6.80 (0.08)	6.80 (0.08)	6.80 (0.08)	6.80 (0.08)	6.80 (0.08)	6.80 (0.08)	6.80 (0.08)	6.80 (0.08)	6.80 (0.08)
$RL_{\text{geo-M}}$	7.21 (0.08)	7.21 (0.08)	7.21 (0.08)	7.21 (0.08)	7.21 (0.08)	7.21 (0.08)	7.21 (0.08)	7.21 (0.08)	7.21 (0.08)	7.21 (0.08)
ViM	11.11 (0.89)	11.17 (0.97)	10.98 (0.62)	11.11 (0.32)	11.06 (0.25)	10.89 (0.52)	10.93 (0.60)	11.07 (0.56)	11.21 (0.53)	11.09 (0.46)
$SIRC$	7.61 (0.07)	7.58 (0.09)	7.56 (0.09)	7.56 (0.07)	7.54 (0.08)	7.53 (0.09)	7.52 (0.09)	7.52 (0.10)	7.53 (0.09)	7.53 (0.09)
KNN	7.16 (0.38)	6.74 (0.10)	6.65 (0.05)	6.67 (0.09)	6.67 (0.09)	6.67 (0.11)	6.67 (0.12)	6.65 (0.13)	6.66 (0.12)	6.66 (0.10)
AURC-MLP	11.16 (0.75)	9.84 (1.01)	8.62 (0.59)	8.76 (0.70)	8.82 (0.82)	8.21 (0.41)	7.85 (0.36)	7.57 (0.44)	7.40 (0.26)	7.14 (0.21)
SELE-SCR	**6.77** (0.13)	6.71 (0.08)	6.66 (0.17)	6.64 (0.07)	6.61 (0.07)	6.64 (0.05)	6.59 (0.08)	6.56 (0.06)	6.56 (0.06)	6.54 (0.07)
SELE-SLR	6.83 (0.22)	**6.68** (0.07)	**6.58** (0.11)	**6.57** (0.06)	**6.56** (0.06)	**6.56** (0.08)	**6.55** (0.09)	**6.54** (0.08)	**6.55** (0.08)	**6.53** (0.08)
AURC-SCR	6.79 (0.09)	**6.63** (0.08)	6.59 (0.10)	6.59 (0.10)	6.60 (0.10)	6.58 (0.06)	6.56 (0.07)	6.58 (0.09)	6.57 (0.06)	6.54 (0.06)
AURC-SLR	**6.74** (0.12)	6.77 (0.21)	**6.56** (0.09)	**6.56** (0.09)	**6.53** (0.11)	**6.55** (0.14)	**6.51** (0.06)	**6.49** (0.06)	**6.49** (0.07)	**6.48** (0.06)

Table 12: Detailed calibration-size sweep on CIFAR100. Entries are mean (standard deviation) over 5 repeats. **Blue**-background rows are our methods. **Green** numbers indicate the best mean score at each calibration size. **Blue** numbers indicate the second-best mean.

	CIFAR100									
Method	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
RL_{max}	28.44 (0.33)	28.44 (0.33)	28.44 (0.33)	28.44 (0.33)	28.44 (0.33)	28.44 (0.33)	28.44 (0.33)	28.44 (0.33)	28.44 (0.33)	28.44 (0.33)
SR_{max}	27.81 (0.29)	27.81 (0.29)	27.81 (0.29)	27.81 (0.29)	27.81 (0.29)	27.81 (0.29)	27.81 (0.29)	27.81 (0.29)	27.81 (0.29)	27.81 (0.29)
SR_{doctor}	27.81 (0.29)	27.81 (0.29)	27.81 (0.29)	27.81 (0.29)	27.81 (0.29)	27.81 (0.29)	27.81 (0.29)	27.81 (0.29)	27.81 (0.29)	27.81 (0.29)
SR_{ent}	27.87 (0.29)	27.87 (0.29)	27.87 (0.29)	27.87 (0.29)	27.87 (0.29)	27.87 (0.29)	27.87 (0.29)	27.87 (0.29)	27.87 (0.29)	27.87 (0.29)
$Energy$	28.43 (0.33)	28.43 (0.33)	28.43 (0.33)	28.43 (0.33)	28.43 (0.33)	28.43 (0.33)	28.43 (0.33)	28.43 (0.33)	28.43 (0.33)	28.43 (0.33)
$RL_{\text{conf-M}}$	25.70 (0.25)	25.70 (0.25)	25.70 (0.25)	25.70 (0.25)	25.70 (0.25)	25.70 (0.25)	25.70 (0.25)	25.70 (0.25)	25.70 (0.25)	25.70 (0.25)
$RL_{\text{geo-M}}$	25.75 (0.28)	25.75 (0.28)	25.75 (0.28)	25.75 (0.28)	25.75 (0.28)	25.75 (0.28)	25.75 (0.28)	25.75 (0.28)	25.75 (0.28)	25.75 (0.28)
ViM	35.41 (3.19)	36.29 (1.81)	36.20 (2.18)	36.51 (1.64)	36.40 (2.14)	36.39 (2.33)	36.45 (1.96)	36.27 (1.37)	36.38 (1.20)	36.71 (1.08)
$SIRC$	28.42 (0.50)	28.55 (0.56)	28.55 (0.45)	28.65 (0.38)	28.59 (0.43)	28.54 (0.42)	28.55 (0.42)	28.55 (0.41)	28.54 (0.41)	28.56 (0.44)
KNN	24.93 (0.83)	**24.09** (0.27)	24.05 (0.24)	23.98 (0.31)	23.96 (0.30)	23.94 (0.28)	23.89 (0.29)	23.88 (0.26)	23.89 (0.26)	23.87 (0.28)
AURC-MLP	29.85 (3.99)	27.26 (0.89)	26.96 (1.30)	26.27 (1.08)	26.25 (1.67)	25.28 (0.56)	24.75 (0.64)	24.60 (0.51)	24.44 (0.40)	24.18 (0.53)
SELE-SCR	25.54 (0.80)	24.58 (0.42)	24.43 (0.34)	24.20 (0.19)	24.11 (0.27)	24.09 (0.17)	24.09 (0.21)	24.04 (0.28)	24.05 (0.29)	24.07 (0.26)
SELE-SLR	**24.76** (0.36)	24.44 (0.42)	24.27 (0.33)	24.10 (0.32)	23.96 (0.29)	24.02 (0.31)	23.92 (0.26)	24.00 (0.24)	23.98 (0.21)	23.91 (0.31)
AURC-SCR	24.95 (0.46)	24.21 (0.22)	**24.00** (0.21)	**23.93** (0.26)	**23.85** (0.25)	**23.82** (0.26)	**23.77** (0.25)	**23.82** (0.21)	**23.82** (0.21)	23.79 (0.24)
AURC-SLR	**24.66** (0.32)	**24.18** (0.21)	**23.99** (0.12)	**23.89** (0.25)	**23.80** (0.28)	**23.78** (0.25)	**23.73** (0.23)	**23.76** (0.19)	**23.77** (0.22)	**23.76** (0.26)

Table 13: Detailed calibration-size sweep on EYEPACS. Entries are mean (standard deviation) over 5 repeats. **Blue**-background rows are our methods. **Green** numbers indicate the best mean score at each calibration size. **Blue** numbers indicate the second-best mean.

	EYEPACS									
Method	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
RL_{max}	11.68 (0.14)	11.68 (0.14)	11.68 (0.14)	11.68 (0.14)	11.68 (0.14)	11.68 (0.14)	11.68 (0.14)	11.68 (0.14)	11.68 (0.14)	11.68 (0.14)
SR_{max}	**9.33** (0.15)	9.33 (0.15)	9.33 (0.15)	9.33 (0.15)	9.33 (0.15)	9.33 (0.15)	9.33 (0.15)	9.33 (0.15)	9.33 (0.15)	9.33 (0.15)
SR_{doctor}	9.36 (0.15)	9.36 (0.15)	9.36 (0.15)	9.36 (0.15)	9.36 (0.15)	9.36 (0.15)	9.36 (0.15)	9.36 (0.15)	9.36 (0.15)	9.36 (0.15)
SR_{ent}	9.44 (0.14)	9.44 (0.14)	9.44 (0.14)	9.44 (0.14)	9.44 (0.14)	9.44 (0.14)	9.44 (0.14)	9.44 (0.14)	9.44 (0.14)	9.44 (0.14)
$Energy$	12.51 (0.15)	12.51 (0.15)	12.51 (0.15)	12.51 (0.15)	12.51 (0.15)	12.51 (0.15)	12.51 (0.15)	12.51 (0.15)	12.51 (0.15)	12.51 (0.15)
$RL_{\text{conf-M}}$	9.75 (0.20)	9.75 (0.20)	9.75 (0.20)	9.75 (0.20)	9.75 (0.20)	9.75 (0.20)	9.75 (0.20)	9.75 (0.20)	9.75 (0.20)	9.75 (0.20)
$RL_{\text{geo-M}}$	15.52 (0.30)	15.52 (0.30)	15.52 (0.30)	15.52 (0.30)	15.52 (0.30)	15.52 (0.30)	15.52 (0.30)	15.52 (0.30)	15.52 (0.30)	15.52 (0.30)
ViM	10.14 (0.17)	10.18 (0.15)	10.19 (0.08)	10.22 (0.10)	10.22 (0.11)	10.23 (0.10)	10.24 (0.08)	10.24 (0.08)	10.25 (0.08)	10.27 (0.07)
$SIRC$	**9.10** (0.14)	**9.12** (0.13)	**9.11** (0.14)	**9.11** (0.14)	**9.11** (0.15)	**9.10** (0.15)	**9.11** (0.15)	**9.10** (0.14)	**9.11** (0.15)	**9.11** (0.15)
KNN	17.27 (1.48)	17.25 (1.18)	17.39 (0.53)	17.46 (0.29)	17.57 (0.18)	17.61 (0.18)	17.67 (0.16)	17.66 (0.25)	17.72 (0.32)	17.77 (0.19)
AURC-MLP	14.63 (2.41)	13.74 (0.81)	13.83 (0.99)	13.75 (2.14)	12.89 (1.10)	13.19 (0.98)	12.79 (0.89)	13.36 (1.14)	13.66 (1.00)	13.20 (1.80)
SELE-SCR	10.09 (1.25)	9.46 (0.13)	9.34 (0.11)	9.29 (0.15)	9.25 (0.14)	9.29 (0.20)	9.24 (0.21)	9.23 (0.22)	9.22 (0.24)	9.20 (0.22)
SELE-SLR	10.07 (1.14)	9.45 (0.13)	9.32 (0.09)	**9.26** (0.14)	9.24 (0.15)	9.27 (0.19)	9.24 (0.22)	9.22 (0.21)	**9.20** (0.22)	9.19 (0.20)
AURC-SCR	10.22 (1.52)	9.51 (0.27)	9.41 (0.10)	9.34 (0.12)	9.34 (0.20)	9.38 (0.19)	9.35 (0.25)	9.27 (0.23)	9.25 (0.22)	9.21 (0.16)
AURC-SLR	9.60 (0.27)	9.42 (0.13)	9.41 (0.15)	9.31 (0.16)	9.31 (0.16)	9.36 (0.20)	9.33 (0.30)	9.22 (0.21)	9.23 (0.22)	9.23 (0.21)

Table 14: Detailed calibration-size sweep on IMAGENET. Entries are mean (standard deviation) over 5 repeats. **Blue**-background rows are our methods. **Green** numbers indicate the best mean score at each calibration size. **Blue** numbers indicate the second-best mean.

	IMAGENET									
Method	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
RL_{max}	4.88 (0.15)	4.88 (0.15)	4.88 (0.15)	4.88 (0.15)	4.88 (0.15)	4.88 (0.15)	4.88 (0.15)	4.88 (0.15)	4.88 (0.15)	4.88 (0.15)
SR_{max}	4.17 (0.12)	4.17 (0.12)	4.17 (0.12)	4.17 (0.12)	4.17 (0.12)	4.17 (0.12)	4.17 (0.12)	4.17 (0.12)	4.17 (0.12)	4.17 (0.12)
SR_{doctor}	4.23 (0.12)	4.23 (0.12)	4.23 (0.12)	4.23 (0.12)	4.23 (0.12)	4.23 (0.12)	4.23 (0.12)	4.23 (0.12)	4.23 (0.12)	4.23 (0.12)
SR_{ent}	5.43 (0.17)	5.43 (0.17)	5.43 (0.17)	5.43 (0.17)	5.43 (0.17)	5.43 (0.17)	5.43 (0.17)	5.43 (0.17)	5.43 (0.17)	5.43 (0.17)
$Energy$	6.24 (0.19)	6.24 (0.19)	6.24 (0.19)	6.24 (0.19)	6.24 (0.19)	6.24 (0.19)	6.24 (0.19)	6.24 (0.19)	6.24 (0.19)	6.24 (0.19)
$RL_{\text{conf-M}}$	**3.25** (0.10)	**3.25** (0.10)	**3.25** (0.10)	**3.25** (0.10)	3.25 (0.10)	3.25 (0.10)	3.25 (0.10)	3.25 (0.10)	3.25 (0.10)	3.25 (0.10)
$RL_{\text{geo-M}}$	**3.27** (0.09)	**3.27** (0.09)	**3.27** (0.09)	3.27 (0.09)	3.27 (0.09)	3.27 (0.09)	3.27 (0.09)	3.27 (0.09)	3.27 (0.09)	3.27 (0.09)
ViM	7.93 (0.14)	8.30 (0.14)	8.49 (0.15)	8.56 (0.24)	8.66 (0.20)	8.69 (0.29)	8.71 (0.32)	8.69 (0.27)	8.68 (0.30)	8.66 (0.32)
$SIRC$	3.98 (0.13)	3.99 (0.12)	4.00 (0.13)	4.00 (0.13)	4.01 (0.13)	4.01 (0.13)	4.01 (0.12)	4.01 (0.12)	4.01 (0.12)	4.01 (0.12)
KNN	20.59 (0.79)	20.24 (0.50)	19.67 (0.62)	18.92 (0.36)	18.24 (0.45)	17.43 (0.59)	16.90 (0.54)	16.34 (0.43)	15.64 (0.51)	15.13 (0.59)
AURC-MLP	12.70 (1.69)	12.00 (0.59)	11.76 (0.90)	11.26 (0.73)	11.25 (1.48)	11.31 (0.82)	10.71 (0.64)	10.68 (0.62)	10.79 (0.34)	10.18 (0.70)
SELE-SCR	3.39 (0.17)	3.33 (0.19)	**3.25** (0.13)	**3.24** (0.11)	**3.24** (0.09)	**3.24** (0.11)	**3.24** (0.13)	**3.23** (0.13)	**3.22** (0.11)	**3.22** (0.10)
SELE-SLR	3.41 (0.21)	3.33 (0.20)	**3.25** (0.13)	**3.24** (0.11)	**3.23** (0.09)	**3.23** (0.11)	**3.23** (0.13)	**3.23** (0.13)	**3.22** (0.11)	**3.22** (0.09)
AURC-SCR	3.53 (0.32)	3.38 (0.30)	3.29 (0.15)	**3.25** (0.11)	3.29 (0.13)	3.26 (0.10)	3.26 (0.13)	**3.24** (0.12)	**3.24** (0.11)	**3.24** (0.09)
AURC-SLR	3.52 (0.33)	3.37 (0.30)	3.28 (0.14)	**3.25** (0.10)	3.28 (0.11)	3.26 (0.10)	3.25 (0.12)	**3.24** (0.13)	**3.24** (0.11)	**3.24** (0.09)

Table 15: Detailed calibration-size sweep on IMAGENET-O. Entries are mean (standard deviation) over 5 repeats. **Blue**-background rows are our methods. **Green** numbers indicate the best mean score at each calibration size. **Blue** numbers indicate the second-best mean.

	IMAGENET-O									
Method	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
RL_{max}	5.94 (0.15)	5.94 (0.15)	5.94 (0.15)	5.94 (0.15)	5.94 (0.15)	5.94 (0.15)	5.94 (0.15)	5.94 (0.15)	5.94 (0.15)	5.94 (0.15)
SR_{max}	5.67 (0.10)	5.67 (0.10)	5.67 (0.10)	5.67 (0.10)	5.67 (0.10)	5.67 (0.10)	5.67 (0.10)	5.67 (0.10)	5.67 (0.10)	5.67 (0.10)
SR_{doctor}	5.71 (0.10)	5.71 (0.10)	5.71 (0.10)	5.71 (0.10)	5.71 (0.10)	5.71 (0.10)	5.71 (0.10)	5.71 (0.10)	5.71 (0.10)	5.71 (0.10)
SR_{ent}	6.85 (0.12)	6.85 (0.12)	6.85 (0.12)	6.85 (0.12)	6.85 (0.12)	6.85 (0.12)	6.85 (0.12)	6.85 (0.12)	6.85 (0.12)	6.85 (0.12)
$Energy$	7.26 (0.22)	7.26 (0.22)	7.26 (0.22)	7.26 (0.22)	7.26 (0.22)	7.26 (0.22)	7.26 (0.22)	7.26 (0.22)	7.26 (0.22)	7.26 (0.22)
$RL_{\text{conf-M}}$	**4.80** (0.08)	4.80 (0.08)	4.80 (0.08)	4.80 (0.08)	4.80 (0.08)	4.80 (0.08)	4.80 (0.08)	4.80 (0.08)	4.80 (0.08)	4.80 (0.08)
$RL_{\text{geo-M}}$	**4.78** (0.09)	4.78 (0.09)	4.78 (0.09)	4.78 (0.09)	4.78 (0.09)	4.78 (0.09)	4.78 (0.09)	4.78 (0.09)	4.78 (0.09)	4.78 (0.09)
ViM	8.72 (0.37)	9.29 (0.34)	9.46 (0.41)	9.57 (0.33)	9.64 (0.30)	9.63 (0.31)	9.62 (0.30)	9.56 (0.33)	9.55 (0.36)	9.51 (0.34)
$SIRC$	5.28 (0.09)	5.29 (0.09)	5.30 (0.08)	5.30 (0.08)	5.30 (0.08)	5.30 (0.09)	5.30 (0.08)	5.30 (0.08)	5.30 (0.08)	5.30 (0.08)
KNN	23.14 (0.79)	22.41 (0.24)	21.60 (0.30)	20.97 (0.35)	20.38 (0.44)	19.53 (0.58)	18.76 (0.44)	18.12 (0.24)	17.48 (0.42)	16.89 (0.51)
AURC-MLP	15.08 (2.07)	13.59 (0.99)	13.44 (1.13)	13.09 (1.30)	13.08 (1.32)	12.42 (1.17)	12.67 (1.04)	11.90 (1.18)	12.68 (0.77)	11.62 (0.87)
SELE-SCR	4.94 (0.35)	**4.70** (0.12)	**4.66** (0.10)	**4.63** (0.12)	**4.60** (0.11)	**4.60** (0.08)	**4.60** (0.08)	4.61 (0.10)	**4.59** (0.07)	4.59 (0.08)
SELE-SLR	4.93 (0.34)	**4.70** (0.12)	**4.66** (0.10)	**4.63** (0.12)	**4.60** (0.11)	**4.61** (0.08)	**4.59** (0.08)	**4.58** (0.08)	**4.59** (0.07)	**4.58** (0.06)
AURC-SCR	4.91 (0.31)	4.81 (0.18)	4.72 (0.14)	4.65 (0.15)	4.62 (0.12)	4.62 (0.10)	4.61 (0.07)	4.60 (0.08)	4.60 (0.08)	**4.58** (0.08)
AURC-SLR	4.91 (0.33)	**4.73** (0.16)	**4.68** (0.13)	**4.61** (0.13)	**4.59** (0.10)	4.62 (0.10)	**4.59** (0.03)	**4.55** (0.06)	**4.58** (0.08)	**4.56** (0.10)

Table 16: Detailed calibration-size sweep on IWILDCAM-ID. Entries are mean (standard deviation) over 5 repeats. **Blue**-background rows are our methods. **Green** numbers indicate the best mean score at each calibration size. **Blue** numbers indicate the second-best mean.

	IWILDCAM-ID									
Method	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
RL_{max}	23.91 (0.30)	23.91 (0.30)	23.91 (0.30)	23.91 (0.30)	23.91 (0.30)	23.91 (0.30)	23.91 (0.30)	23.91 (0.30)	23.91 (0.30)	23.91 (0.30)
SR_{max}	9.82 (0.39)	9.82 (0.39)	9.82 (0.39)	9.82 (0.39)	9.82 (0.39)	9.82 (0.39)	9.82 (0.39)	9.82 (0.39)	9.82 (0.39)	9.82 (0.39)
SR_{doctor}	9.92 (0.39)	9.92 (0.39)	9.92 (0.39)	9.92 (0.39)	9.92 (0.39)	9.92 (0.39)	9.92 (0.39)	9.92 (0.39)	9.92 (0.39)	9.92 (0.39)
SR_{ent}	10.34 (0.39)	10.34 (0.39)	10.34 (0.39)	10.34 (0.39)	10.34 (0.39)	10.34 (0.39)	10.34 (0.39)	10.34 (0.39)	10.34 (0.39)	10.34 (0.39)
$Energy$	28.95 (0.41)	28.95 (0.41)	28.95 (0.41)	28.95 (0.41)	28.95 (0.41)	28.95 (0.41)	28.95 (0.41)	28.95 (0.41)	28.95 (0.41)	28.95 (0.41)
$RL_{\text{conf-M}}$	9.97 (0.44)	9.97 (0.44)	9.97 (0.44)	9.97 (0.44)	9.97 (0.44)	9.97 (0.44)	9.97 (0.44)	9.97 (0.44)	9.97 (0.44)	9.97 (0.44)
$RL_{\text{geo-M}}$	9.98 (0.44)	9.98 (0.44)	9.98 (0.44)	9.98 (0.44)	9.98 (0.44)	9.98 (0.44)	9.98 (0.44)	9.98 (0.44)	9.98 (0.44)	9.98 (0.44)
ViM	14.71 (1.39)	16.17 (1.16)	18.03 (1.41)	18.02 (1.36)	17.81 (0.97)	17.61 (1.45)	16.99 (1.66)	16.68 (1.53)	16.57 (1.37)	16.66 (1.50)
$SIRC$	9.63 (0.39)	9.63 (0.40)	9.65 (0.40)	9.65 (0.40)	9.68 (0.37)	9.68 (0.37)	9.69 (0.38)	9.68 (0.37)	9.69 (0.37)	9.70 (0.36)
KNN	12.36 (1.04)	11.69 (1.36)	11.98 (1.47)	12.47 (1.20)	13.23 (0.59)	13.74 (0.60)	14.03 (0.70)	14.62 (1.09)	15.33 (0.69)	15.76 (0.87)
AURC-MLP	10.81 (0.49)	9.80 (0.82)	8.87 (0.66)	9.09 (0.63)	**8.70** (0.32)	**8.61** (0.27)	**8.44** (0.47)	**7.89** (0.21)	**7.84** (0.39)	**7.58** (0.52)
SELE-SCR	8.68 (0.31)	8.54 (0.48)	8.68 (0.52)	8.82 (0.57)	9.04 (0.49)	9.08 (0.37)	9.12 (0.38)	9.20 (0.43)	9.34 (0.35)	9.43 (0.38)
SELE-SLR	**8.62** (0.33)	8.53 (0.49)	8.68 (0.54)	8.81 (0.49)	9.01 (0.48)	9.07 (0.38)	9.13 (0.39)	9.22 (0.50)	9.36 (0.40)	9.40 (0.41)
AURC-SCR	8.69 (0.35)	8.49 (0.60)	8.65 (0.50)	**8.80** (0.60)	8.93 (0.42)	8.97 (0.24)	9.04 (0.19)	9.14 (0.39)	9.26 (0.29)	9.32 (0.34)
AURC-SLR	**8.32** (0.46)	**8.37** (0.46)	**8.42** (0.48)	**8.61** (0.44)	**8.81** (0.33)	**8.80** (0.27)	**8.76** (0.24)	**8.92** (0.37)	**9.03** (0.30)	**9.08** (0.46)

Table 17: Detailed calibration-size sweep on IWILDCAM-OOD. Entries are mean (standard deviation) over 5 repeats. **Blue**-background rows are our methods. **Green** numbers indicate the best mean score at each calibration size. **Blue** numbers indicate the second-best mean.

	IWILDCAM-OOD									
Method	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
RL_{max}	28.61 (0.39)	28.61 (0.39)	28.61 (0.39)	28.61 (0.39)	28.61 (0.39)	28.61 (0.39)	28.61 (0.39)	28.61 (0.39)	28.61 (0.39)	28.61 (0.39)
SR_{max}	10.20 (0.10)	10.20 (0.10)	10.20 (0.10)	10.20 (0.10)	10.20 (0.10)	10.20 (0.10)	10.20 (0.10)	10.20 (0.10)	10.20 (0.10)	10.20 (0.10)
SR_{doctor}	10.22 (0.10)	10.22 (0.10)	10.22 (0.10)	10.22 (0.10)	10.22 (0.10)	10.22 (0.10)	10.22 (0.10)	10.22 (0.10)	10.22 (0.10)	10.22 (0.10)
SR_{ent}	10.66 (0.09)	10.66 (0.09)	10.66 (0.09)	10.66 (0.09)	10.66 (0.09)	10.66 (0.09)	10.66 (0.09)	10.66 (0.09)	10.66 (0.09)	10.66 (0.09)
$Energy$	35.38 (0.51)	35.38 (0.51)	35.38 (0.51)	35.38 (0.51)	35.38 (0.51)	35.38 (0.51)	35.38 (0.51)	35.38 (0.51)	35.38 (0.51)	35.38 (0.51)
$RL_{\text{conf-M}}$	10.12 (0.13)	10.12 (0.13)	10.12 (0.13)	10.12 (0.13)	10.12 (0.13)	10.12 (0.13)	10.12 (0.13)	10.12 (0.13)	10.12 (0.13)	10.12 (0.13)
$RL_{\text{geo-M}}$	10.12 (0.13)	10.12 (0.13)	10.12 (0.13)	10.12 (0.13)	10.12 (0.13)	10.12 (0.13)	10.12 (0.13)	10.12 (0.13)	10.12 (0.13)	10.12 (0.13)
ViM	11.96 (0.32)	12.22 (0.54)	12.37 (0.26)	12.31 (0.20)	12.34 (0.20)	12.37 (0.32)	12.30 (0.24)	12.21 (0.24)	12.18 (0.24)	12.13 (0.30)
$SIRC$	10.04 (0.12)	10.03 (0.11)	10.01 (0.09)	10.01 (0.09)	10.01 (0.09)	10.01 (0.09)	10.01 (0.09)	10.01 (0.09)	10.01 (0.09)	10.02 (0.09)
KNN	12.15 (0.63)	11.16 (0.26)	10.76 (0.18)	10.67 (0.23)	10.65 (0.26)	10.73 (0.32)	10.67 (0.25)	10.61 (0.32)	10.63 (0.36)	10.67 (0.36)
AURC-MLP	9.67 (0.17)	9.52 (0.65)	9.67 (0.17)	9.28 (0.38)	8.95 (0.38)	8.74 (0.28)	8.58 (0.28)	8.50 (0.24)	8.46 (0.25)	**8.28** (0.24)
SELE-SCR	8.89 (0.27)	8.83 (0.15)	8.65 (0.16)	8.63 (0.24)	8.66 (0.23)	8.61 (0.12)	8.51 (0.09)	8.50 (0.04)	8.53 (0.08)	8.57 (0.06)
SELE-SLR	8.84 (0.23)	8.55 (0.47)	8.43 (0.12)	8.48 (0.26)	8.48 (0.15)	8.54 (0.11)	8.48 (0.18)	8.40 (0.22)	8.44 (0.19)	8.49 (0.17)
AURC-SCR	8.82 (0.16)	8.70 (0.07)	8.50 (0.08)	**8.43** (0.15)	**8.38** (0.10)	**8.36** (0.10)	**8.34** (0.07)	**8.30** (0.10)	**8.33** (0.10)	8.36 (0.09)
AURC-SLR	**8.48** (0.20)	**8.27** (0.08)	**8.17** (0.17)	**8.22** (0.16)	**8.19** (0.07)	**8.15** (0.07)	**8.14** (0.08)	**8.14** (0.10)	**8.16** (0.09)	**8.16** (0.08)

Table 18: Detailed calibration-size sweep on OPENIMAGE-O. Entries are mean (standard deviation) over 5 repeats. **Blue**-background rows are our methods. **Green** numbers indicate the best mean score at each calibration size. **Blue** numbers indicate the second-best mean.

	OPENIMAGE-O									
Method	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
RL_{max}	**12.02** (0.13)	12.02 (0.13)	12.02 (0.13)	12.02 (0.13)	12.02 (0.13)	12.02 (0.13)	12.02 (0.13)	12.02 (0.13)	12.02 (0.13)	12.02 (0.13)
SR_{max}	12.91 (0.08)	12.91 (0.08)	12.91 (0.08)	12.91 (0.08)	12.91 (0.08)	12.91 (0.08)	12.91 (0.08)	12.91 (0.08)	12.91 (0.08)	12.91 (0.08)
SR_{doctor}	12.86 (0.08)	12.86 (0.08)	12.86 (0.08)	12.86 (0.08)	12.86 (0.08)	12.86 (0.08)	12.86 (0.08)	12.86 (0.08)	12.86 (0.08)	12.86 (0.08)
SR_{ent}	13.96 (0.14)	13.96 (0.14)	13.96 (0.14)	13.96 (0.14)	13.96 (0.14)	13.96 (0.14)	13.96 (0.14)	13.96 (0.14)	13.96 (0.14)	13.96 (0.14)
$Energy$	13.28 (0.14)	13.28 (0.14)	13.28 (0.14)	13.28 (0.14)	13.28 (0.14)	13.28 (0.14)	13.28 (0.14)	13.28 (0.14)	13.28 (0.14)	13.28 (0.14)
$RL_{\text{conf-M}}$	12.84 (0.13)	12.84 (0.13)	12.84 (0.13)	12.84 (0.13)	12.84 (0.13)	12.84 (0.13)	12.84 (0.13)	12.84 (0.13)	12.84 (0.13)	12.84 (0.13)
$RL_{\text{geo-M}}$	12.77 (0.12)	12.77 (0.12)	12.77 (0.12)	12.77 (0.12)	12.77 (0.12)	12.77 (0.12)	12.77 (0.12)	12.77 (0.12)	12.77 (0.12)	12.77 (0.12)
ViM	16.19 (0.38)	17.30 (0.35)	18.05 (0.25)	18.36 (0.30)	18.63 (0.44)	18.98 (0.56)	19.06 (0.55)	19.16 (0.56)	19.07 (0.53)	18.86 (0.46)
$SIRC$	**12.28** (0.05)	12.32 (0.07)	12.35 (0.07)	12.36 (0.07)	12.38 (0.07)	12.38 (0.07)	12.39 (0.07)	12.40 (0.07)	12.39 (0.07)	12.39 (0.07)
KNN	56.82 (0.54)	54.90 (1.03)	53.34 (1.50)	51.45 (1.34)	49.87 (0.92)	48.49 (1.37)	46.98 (1.43)	45.37 (1.19)	43.73 (1.19)	42.30 (1.23)
AURC-MLP	16.43 (0.74)	15.45 (0.55)	15.26 (0.63)	15.25 (0.62)	15.35 (0.65)	15.01 (0.47)	15.08 (0.51)	15.01 (0.47)	15.15 (0.69)	15.23 (0.41)
SELE-SCR	12.81 (1.03)	**11.31** (0.15)	11.37 (0.19)	11.30 (0.11)	11.31 (0.17)	11.30 (0.15)	11.31 (0.17)	11.28 (0.14)	11.29 (0.14)	11.28 (0.14)
SELE-SLR	12.76 (1.01)	**11.32** (0.16)	**11.36** (0.15)	11.28 (0.12)	**11.28** (0.20)	**11.23** (0.24)	**11.26** (0.17)	**11.21** (0.13)	**11.18** (0.13)	**11.20** (0.11)
AURC-SCR	12.73 (1.93)	11.41 (0.36)	11.42 (0.15)	**11.27** (0.16)	11.29 (0.19)	11.26 (0.13)	11.27 (0.16)	11.23 (0.12)	11.24 (0.12)	11.24 (0.13)
AURC-SLR	12.37 (1.35)	11.39 (0.35)	**11.32** (0.10)	**11.17** (0.18)	**11.20** (0.24)	**11.12** (0.09)	**11.09** (0.13)	**11.06** (0.11)	**11.05** (0.14)	**11.07** (0.10)

Table 19: Detailed calibration-size sweep on NIHLT. Entries are mean (standard deviation) over 5 repeats. **Blue**-background rows are our methods. **Green** numbers indicate the best mean score at each calibration size. **Blue** numbers indicate the second-best mean.

	NIHLT									
Method	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
RL_{max}	77.05 (0.71)	77.05 (0.71)	77.05 (0.71)	77.05 (0.71)	77.05 (0.71)	77.05 (0.71)	77.05 (0.71)	77.05 (0.71)	77.05 (0.71)	77.05 (0.71)
SR_{max}	78.30 (0.59)	78.30 (0.59)	78.30 (0.59)	78.30 (0.59)	78.30 (0.59)	78.30 (0.59)	78.30 (0.59)	78.30 (0.59)	78.30 (0.59)	78.30 (0.59)
SR_{doctor}	77.65 (0.61)	77.65 (0.61)	77.65 (0.61)	77.65 (0.61)	77.65 (0.61)	77.65 (0.61)	77.65 (0.61)	77.65 (0.61)	77.65 (0.61)	77.65 (0.61)
SR_{ent}	77.61 (0.60)	77.61 (0.60)	77.61 (0.60)	77.61 (0.60)	77.61 (0.60)	77.61 (0.60)	77.61 (0.60)	77.61 (0.60)	77.61 (0.60)	77.61 (0.60)
$Energy$	74.26 (0.89)	74.26 (0.89)	74.26 (0.89)	74.26 (0.89)	74.26 (0.89)	74.26 (0.89)	74.26 (0.89)	74.26 (0.89)	74.26 (0.89)	74.26 (0.89)
$RL_{\text{conf-M}}$	78.25 (0.70)	78.25 (0.70)	78.25 (0.70)	78.25 (0.70)	78.25 (0.70)	78.25 (0.70)	78.25 (0.70)	78.25 (0.70)	78.25 (0.70)	78.25 (0.70)
$RL_{\text{geo-M}}$	77.73 (0.72)	77.73 (0.72)	77.73 (0.72)	77.73 (0.72)	77.73 (0.72)	77.73 (0.72)	77.73 (0.72)	77.73 (0.72)	77.73 (0.72)	77.73 (0.72)
ViM	76.16 (0.83)	75.58 (0.74)	75.52 (0.70)	75.57 (0.70)	75.48 (0.73)	75.34 (0.71)	75.39 (0.75)	75.38 (0.75)	75.39 (0.75)	75.39 (0.74)
$SIRC$	78.47 (0.59)	78.08 (0.50)	78.05 (0.51)	78.05 (0.56)	77.95 (0.67)	77.85 (0.59)	77.88 (0.65)	77.90 (0.65)	77.91 (0.64)	77.91 (0.65)
KNN	73.79 (0.84)	73.37 (0.77)	73.31 (0.60)	73.04 (0.79)	72.99 (0.88)	72.77 (1.06)	**72.58** (0.99)	72.66 (1.01)	**72.59** (0.93)	72.55 (0.90)
AURC-MLP	**72.22** (1.82)	**71.66** (2.27)	**71.39** (1.28)	**71.19** (1.30)	**70.55** (1.29)	**70.08** (1.24)	**70.10** (1.11)	**69.54** (1.24)	**70.87** (4.04)	**71.12** (4.16)
SELE-SCR	73.88 (1.00)	73.47 (1.41)	72.98 (0.65)	73.37 (1.60)	72.94 (1.07)	72.92 (1.28)	72.87 (1.15)	72.85 (1.13)	72.84 (1.19)	72.47 (0.99)
SELE-SLR	73.68 (0.95)	73.47 (1.37)	73.00 (0.51)	73.29 (1.55)	72.94 (1.10)	72.83 (1.29)	72.79 (1.14)	72.68 (1.04)	72.65 (1.13)	72.39 (1.04)
AURC-SCR	73.66 (1.02)	73.36 (1.47)	72.96 (0.61)	73.29 (1.59)	72.93 (1.05)	72.77 (1.27)	72.78 (1.13)	72.68 (1.06)	72.80 (1.13)	72.45 (0.93)
AURC-SLR	**73.44** (1.07)	**73.32** (1.49)	**72.91** (0.66)	**72.90** (1.23)	**72.73** (1.01)	**72.66** (1.32)	72.60 (1.14)	**72.58** (1.24)	72.69 (1.08)	**72.25** (0.98)

Table 20: Detailed calibration-size sweep on PCAM. Entries are mean (standard deviation) over 5 repeats. **Blue**-background rows are our methods. **Green** numbers indicate the best mean score at each calibration size. **Blue** numbers indicate the second-best mean.

	PCAM									
Method	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
RL_{max}	15.70 (0.26)	15.70 (0.26)	15.70 (0.26)	15.70 (0.26)	15.70 (0.26)	15.70 (0.26)	15.70 (0.26)	15.70 (0.26)	15.70 (0.26)	15.70 (0.26)
SR_{max}	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)
SR_{doctor}	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)
SR_{ent}	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)
$Energy$	16.37 (0.25)	16.37 (0.25)	16.37 (0.25)	16.37 (0.25)	16.37 (0.25)	16.37 (0.25)	16.37 (0.25)	16.37 (0.25)	16.37 (0.25)	16.37 (0.25)
$RL_{\text{conf-M}}$	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)
$RL_{\text{geo-M}}$	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)	9.05 (0.28)
ViM	11.41 (0.67)	11.50 (0.54)	11.38 (0.48)	11.36 (0.42)	11.30 (0.38)	11.28 (0.39)	11.26 (0.34)	11.25 (0.34)	11.23 (0.31)	11.22 (0.29)
$SIRC$	9.12 (0.29)	9.12 (0.29)	9.12 (0.29)	9.12 (0.29)	9.12 (0.29)	9.12 (0.29)	9.12 (0.29)	9.12 (0.29)	9.12 (0.28)	9.12 (0.29)
KNN	8.00 (0.95)	8.68 (1.03)	9.48 (0.54)	10.06 (0.46)	10.18 (0.85)	10.33 (0.52)	10.63 (0.81)	10.86 (0.80)	10.89 (0.87)	11.17 (0.80)
AURC-MLP	**5.41** (0.98)	**4.75** (0.65)	**4.50** (0.18)	**4.20** (0.28)	**4.08** (0.13)	**4.03** (0.28)	**4.06** (0.22)	**3.92** (0.32)	**3.69** (0.16)	**3.72** (0.18)
SELE-SCR	8.28 (1.67)	8.25 (1.54)	8.26 (0.87)	8.29 (0.42)	8.18 (0.30)	8.30 (0.14)	8.30 (0.11)	8.44 (0.20)	8.42 (0.27)	8.43 (0.25)
SELE-SLR	7.64 (0.98)	7.53 (0.79)	**6.89** (0.76)	**6.81** (0.15)	**6.76** (0.13)	**6.67** (0.16)	6.61 (0.21)	**6.58** (0.21)	**6.48** (0.18)	6.46 (0.16)
AURC-SCR	7.77 (1.47)	7.97 (1.36)	8.16 (0.72)	8.20 (0.38)	8.13 (0.29)	8.26 (0.09)	8.17 (0.03)	8.26 (0.17)	8.29 (0.19)	8.33 (0.19)
AURC-SLR	**7.03** (0.62)	**7.23** (0.74)	7.20 (0.36)	6.84 (0.25)	6.79 (0.17)	6.72 (0.14)	**6.60** (0.31)	6.61 (0.42)	6.51 (0.25)	**6.45** (0.22)

Table 21: Detailed calibration-size sweep on SST2. Entries are mean (standard deviation) over 5 repeats. **Blue**-background rows are our methods. **Green** numbers indicate the best mean score at each calibration size. **Blue** numbers indicate the second-best mean.

	SST2									
Method	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
RL_{max}	13.11 (1.29)	13.11 (1.29)	13.11 (1.29)	13.11 (1.29)	13.11 (1.29)	13.11 (1.29)	13.11 (1.29)	13.11 (1.29)	13.11 (1.29)	13.11 (1.29)
SR_{max}	**9.07** (0.19)	**9.07** (0.19)	**9.07** (0.19)	**9.07** (0.19)	**9.07** (0.19)	**9.07** (0.19)	9.07 (0.19)	9.07 (0.19)	**9.07** (0.19)	**9.07** (0.19)
SR_{doctor}	**9.07** (0.19)	**9.07** (0.19)	**9.07** (0.19)	**9.07** (0.19)	**9.07** (0.19)	**9.07** (0.19)	9.07 (0.19)	9.07 (0.19)	**9.07** (0.19)	**9.07** (0.19)
SR_{ent}	**9.07** (0.19)	**9.07** (0.19)	**9.07** (0.19)	**9.07** (0.19)	**9.07** (0.19)	**9.07** (0.19)	9.07 (0.19)	9.07 (0.19)	**9.07** (0.19)	**9.07** (0.19)
$Energy$	13.45 (1.33)	13.45 (1.33)	13.45 (1.33)	13.45 (1.33)	13.45 (1.33)	13.45 (1.33)	13.45 (1.33)	13.45 (1.33)	13.45 (1.33)	13.45 (1.33)
$RL_{\text{conf-M}}$	**9.07** (0.19)	**9.07** (0.19)	**9.07** (0.19)	**9.07** (0.19)	**9.07** (0.19)	**9.07** (0.19)	9.07 (0.19)	9.07 (0.19)	**9.07** (0.19)	**9.07** (0.19)
$RL_{\text{geo-M}}$	**9.07** (0.19)	**9.07** (0.19)	**9.07** (0.19)	**9.07** (0.19)	**9.07** (0.19)	**9.07** (0.19)	9.07 (0.19)	9.07 (0.19)	**9.07** (0.19)	**9.07** (0.19)
ViM	14.89 (0.80)	14.69 (0.85)	14.72 (0.76)	14.71 (0.64)	14.67 (0.59)	14.66 (0.58)	14.66 (0.58)	14.67 (0.59)	14.66 (0.57)	14.66 (0.57)
$SIRC$	**9.06** (0.21)	**9.06** (0.21)	**9.06** (0.21)	**9.06** (0.21)	**9.06** (0.21)	**9.06** (0.21)	**9.06** (0.21)	**9.06** (0.21)	**9.06** (0.21)	**9.06** (0.21)
KNN	12.18 (1.19)	12.28 (1.20)	12.36 (1.06)	12.27 (1.01)	12.30 (1.02)	12.20 (1.03)	12.15 (1.09)	12.14 (1.01)	12.13 (1.00)	12.12 (0.99)
AURC-MLP	19.01 (1.58)	18.99 (2.30)	18.88 (2.04)	19.05 (2.58)	19.01 (2.54)	18.63 (2.05)	18.59 (2.09)	19.17 (1.90)	19.10 (1.70)	18.55 (1.60)
SELE-SCR	9.42 (0.58)	9.28 (0.47)	9.19 (0.26)	9.15 (0.28)	9.11 (0.26)	9.12 (0.26)	9.12 (0.25)	9.10 (0.25)	9.11 (0.25)	9.11 (0.24)
SELE-SLR	9.32 (0.51)	9.24 (0.45)	9.12 (0.23)	9.13 (0.28)	9.10 (0.27)	9.10 (0.25)	9.10 (0.24)	9.09 (0.26)	9.11 (0.25)	9.10 (0.24)
AURC-SCR	9.46 (0.65)	9.35 (0.46)	9.25 (0.37)	9.29 (0.27)	9.20 (0.31)	9.15 (0.22)	**9.04** (0.22)	**9.01** (0.18)	**9.06** (0.19)	9.08 (0.18)
AURC-SLR	9.26 (0.31)	9.27 (0.31)	9.23 (0.28)	9.27 (0.32)	9.20 (0.31)	9.15 (0.22)	**9.04** (0.21)	**9.01** (0.18)	**9.06** (0.19)	**9.07** (0.17)

Table 22: Detailed LLM calibration-size sweep on MMLU. **Blue**-background rows are our methods. **Green** text marks best or tied-best scores, and **Blue** text marks second-best or tied second-best scores, both at displayed precision.

Mistral-7B-v0.1

Method	100	200	300	400	500	600	700	800	900	1000
$RL_{\max}$	**18.56** (0.59)	18.56 (0.59)	18.56 (0.59)	18.56 (0.59)	18.56 (0.59)	18.56 (0.59)	18.56 (0.59)	18.56 (0.59)	18.56 (0.59)	18.56 (0.59)
$SR_{\max}$	**18.54** (0.59)	18.54 (0.59)	18.54 (0.59)	18.54 (0.59)	18.54 (0.59)	18.54 (0.59)	18.54 (0.59)	18.54 (0.59)	18.54 (0.59)	18.54 (0.59)
SR_{doctor}	19.17 (0.57)	19.17 (0.57)	19.17 (0.57)	19.17 (0.57)	19.17 (0.57)	19.17 (0.57)	19.17 (0.57)	19.17 (0.57)	19.17 (0.57)	19.17 (0.57)
SR_{ent}	19.22 (0.57)	19.22 (0.57)	19.22 (0.57)	19.22 (0.57)	19.22 (0.57)	19.22 (0.57)	19.22 (0.57)	19.22 (0.57)	19.22 (0.57)	19.22 (0.57)
$Energy$	19.29 (0.58)	19.29 (0.58)	19.29 (0.58)	19.29 (0.58)	19.29 (0.58)	19.29 (0.58)	19.29 (0.58)	19.29 (0.58)	19.29 (0.58)	19.29 (0.58)
$RL_{\text{conf-M}}$	18.66 (0.64)	18.66 (0.64)	18.66 (0.64)	18.66 (0.64)	18.66 (0.64)	18.66 (0.64)	18.66 (0.64)	18.66 (0.64)	18.66 (0.64)	18.66 (0.64)
$RL_{\text{geo-M}}$	18.76 (0.51)	18.76 (0.51)	18.76 (0.51)	18.76 (0.51)	18.76 (0.51)	18.76 (0.51)	18.76 (0.51)	18.76 (0.51)	18.76 (0.51)	18.76 (0.51)
ViM	54.48 (1.69)	55.83 (1.25)	55.98 (1.02)	56.10 (1.06)	56.12 (1.05)	56.12 (1.08)	56.12 (1.08)	56.11 (1.09)	56.11 (1.11)	56.11 (1.10)
$SIRC$	31.50 (1.11)	32.11 (1.04)	32.15 (0.97)	32.11 (0.86)	32.13 (0.89)	32.12 (0.88)	32.13 (0.92)	32.11 (0.87)	32.12 (0.89)	32.10 (0.89)
KNN	55.10 (1.94)	55.75 (1.16)	55.84 (1.09)	55.80 (1.04)	55.93 (1.05)	55.93 (1.05)	55.91 (1.00)	55.85 (1.04)	55.85 (1.03)	55.85 (1.03)
AURC-MLP	24.89 (2.51)	28.02 (4.59)	26.51 (3.75)	26.41 (4.56)	26.93 (6.49)	24.81 (5.27)	24.22 (3.05)	21.77 (1.49)	23.49 (2.21)	21.16 (1.05)
SELE-SCR	18.65 (0.60)	**18.44** (0.49)	18.42 (0.49)	**18.37** (0.53)	**18.45** (0.53)	**18.39** (0.57)	**18.36** (0.55)	**18.32** (0.55)	**18.30** (0.56)	**18.30** (0.56)
SELE-SLR	18.67 (0.59)	18.45 (0.50)	18.44 (0.47)	**18.37** (0.51)	**18.46** (0.55)	**18.40** (0.60)	**18.36** (0.57)	**18.32** (0.56)	**18.30** (0.56)	**18.31** (0.56)
AURC-SCR	18.72 (0.62)	**18.44** (0.51)	**18.41** (0.56)	**18.36** (0.59)	18.49 (0.58)	18.42 (0.55)	**18.37** (0.55)	18.33 (0.54)	**18.33** (0.55)	18.32 (0.55)
AURC-SLR	18.77 (0.67)	**18.36** (0.55)	**18.37** (0.54)	**18.37** (0.58)	18.47 (0.54)	18.41 (0.55)	18.38 (0.55)	**18.31** (0.55)	**18.33** (0.54)	18.33 (0.56)

Qwen2.5-7B

Method	100	200	300	400	500	600	700	800	900	1000
$RL_{\max}$	10.53 (0.44)	10.53 (0.44)	10.53 (0.44)	10.53 (0.44)	10.53 (0.44)	10.53 (0.44)	10.53 (0.44)	10.53 (0.44)	10.53 (0.44)	10.53 (0.44)
$SR_{\max}$	10.48 (0.44)	10.48 (0.44)	10.48 (0.44)	10.48 (0.44)	10.48 (0.44)	10.48 (0.44)	10.48 (0.44)	10.48 (0.44)	10.48 (0.44)	10.48 (0.44)
SR_{doctor}	11.76 (0.46)	11.76 (0.46)	11.76 (0.46)	11.76 (0.46)	11.76 (0.46)	11.76 (0.46)	11.76 (0.46)	11.76 (0.46)	11.76 (0.46)	11.76 (0.46)
SR_{ent}	11.86 (0.46)	11.86 (0.46)	11.86 (0.46)	11.86 (0.46)	11.86 (0.46)	11.86 (0.46)	11.86 (0.46)	11.86 (0.46)	11.86 (0.46)	11.86 (0.46)
$Energy$	11.97 (0.46)	11.97 (0.46)	11.97 (0.46)	11.97 (0.46)	11.97 (0.46)	11.97 (0.46)	11.97 (0.46)	11.97 (0.46)	11.97 (0.46)	11.97 (0.46)
$RL_{\text{conf-M}}$	**10.32** (0.43)	**10.32** (0.43)	**10.32** (0.43)	**10.32** (0.43)	**10.32** (0.43)	10.32 (0.43)	10.32 (0.43)	10.32 (0.43)	10.32 (0.43)	10.32 (0.43)
$RL_{\text{geo-M}}$	11.21 (0.42)	11.21 (0.42)	11.21 (0.42)	11.21 (0.42)	11.21 (0.42)	11.21 (0.42)	11.21 (0.42)	11.21 (0.42)	11.21 (0.42)	11.21 (0.42)
ViM	40.38 (0.84)	40.70 (1.07)	40.03 (0.76)	40.26 (0.98)	40.31 (0.92)	40.25 (0.92)	40.15 (0.89)	40.14 (0.84)	40.10 (0.88)	40.10 (0.84)
$SIRC$	21.09 (1.00)	21.30 (0.78)	20.88 (0.29)	21.08 (0.53)	21.08 (0.57)	21.00 (0.67)	20.90 (0.68)	20.84 (0.67)	20.75 (0.62)	20.73 (0.62)
KNN	42.21 (1.85)	42.92 (0.71)	42.92 (0.57)	42.89 (0.64)	42.94 (0.66)	42.78 (0.51)	42.70 (0.46)	42.68 (0.50)	42.68 (0.57)	42.64 (0.60)
AURC-MLP	17.93 (2.44)	19.59 (3.07)	16.10 (3.22)	17.07 (4.09)	17.28 (4.47)	16.75 (4.32)	17.31 (3.71)	16.51 (4.44)	16.07 (4.26)	15.77 (3.78)
SELE-SCR	**10.46** (0.58)	10.39 (0.53)	10.42 (0.40)	**10.35** (0.47)	**10.31** (0.52)	**10.26** (0.46)	**10.27** (0.45)	**10.27** (0.47)	**10.26** (0.44)	**10.28** (0.46)
SELE-SLR	**10.46** (0.55)	**10.34** (0.44)	10.43 (0.40)	10.36 (0.47)	**10.31** (0.52)	**10.26** (0.46)	**10.28** (0.46)	**10.26** (0.46)	**10.26** (0.46)	**10.28** (0.45)
AURC-SCR	10.50 (0.58)	10.40 (0.46)	**10.35** (0.44)	10.37 (0.49)	10.35 (0.54)	**10.27** (0.44)	**10.27** (0.44)	**10.27** (0.46)	**10.24** (0.45)	**10.26** (0.46)
AURC-SLR	10.53 (0.56)	10.46 (0.50)	10.45 (0.42)	10.41 (0.48)	10.35 (0.54)	10.28 (0.45)	10.30 (0.45)	**10.27** (0.45)	**10.26** (0.45)	10.30 (0.45)

and seed grid in parallel used approximately 1,500 CPU-core-hours in total. Including preliminary runs, abandoned loss variants, and debugging runs that are not reported in the paper, we estimate that the full research project consumed roughly 2–3 times the compute used for the final reported experiments.

E Existing Assets and Licenses

We use existing datasets, fixed pretrained models, and published post-hoc confidence-score definitions only for benchmarking. The supplementary package does not redistribute third-party datasets, pretrained/fixed model weights, or third-party benchmark files; it provides reproduction code, scripts, and pointers to the original sources. The public license and terms metadata below were checked on May 4, 2026. When an asset does not provide a standalone license in public metadata, we cite the original source, follow the provider access terms, and do not redistribute the asset.

Datasets and benchmarks.

ImageNet. Used under the ImageNet access terms for non-commercial research and educational use; ImageNet notes that it does not own the image copyrights. Source URL: `https://image-net.org/download-images/download-faq`.

ImageNet-O. Released with the Natural Adversarial Examples benchmark under the MIT license. Source URL: `https://github.com/hendrycks/natural-adv-examples`.

OpenImage-O. Derived from Open Images and the OpenOOD benchmark. OpenOOD is MIT-licensed; Open Images annotations are CC BY 4.0, the repository code is Apache-2.0, and the listed images are described as CC BY 2.0 with the Open Images license-status

Table 23: Detailed LLM calibration-size sweep on MMLU-Pro. **Blue**-background rows are our methods. **Green** text marks best or tied-best scores, and **Blue** text marks second-best or tied second-best scores, both at displayed precision.

Mistral-7B-v0.1

Method	100	200	300	400	500	600	700	800	900	1000
$RL_{\max}$	51.91 (1.30)	51.91 (1.30)	51.91 (1.30)	51.91 (1.30)	51.91 (1.30)	51.91 (1.30)	51.91 (1.30)	51.91 (1.30)	51.91 (1.30)	51.91 (1.30)
$SR_{\max}$	**51.88** (1.30)	**51.88** (1.30)	**51.88** (1.30)	**51.88** (1.30)	**51.88** (1.30)	**51.88** (1.30)	**51.88** (1.30)	**51.88** (1.30)	**51.88** (1.30)	51.88 (1.30)
SR_{doctor}	53.06 (1.31)	53.06 (1.31)	53.06 (1.31)	53.06 (1.31)	53.06 (1.31)	53.06 (1.31)	53.06 (1.31)	53.06 (1.31)	53.06 (1.31)	53.06 (1.31)
SR_{ent}	53.17 (1.31)	53.17 (1.31)	53.17 (1.31)	53.17 (1.31)	53.17 (1.31)	53.17 (1.31)	53.17 (1.31)	53.17 (1.31)	53.17 (1.31)	53.17 (1.31)
$Energy$	53.29 (1.31)	53.29 (1.31)	53.29 (1.31)	53.29 (1.31)	53.29 (1.31)	53.29 (1.31)	53.29 (1.31)	53.29 (1.31)	53.29 (1.31)	53.29 (1.31)
$RL_{\text{conf-M}}$	**51.87** (1.03)	**51.87** (1.03)	**51.87** (1.03)	**51.87** (1.03)	**51.87** (1.03)	**51.87** (1.03)	**51.87** (1.03)	**51.87** (1.03)	**51.87** (1.03)	**51.87** (1.03)
$RL_{\text{geo-M}}$	56.47 (0.93)	56.47 (0.93)	56.47 (0.93)	56.47 (0.93)	56.47 (0.93)	56.47 (0.93)	56.47 (0.93)	56.47 (0.93)	56.47 (0.93)	56.47 (0.93)
ViM	77.99 (0.61)	78.08 (0.54)	78.05 (0.59)	78.04 (0.56)	78.06 (0.59)	78.02 (0.56)	78.00 (0.55)	78.01 (0.55)	78.00 (0.52)	78.00 (0.53)
$SIRC$	70.74 (1.48)	70.30 (0.59)	70.07 (0.40)	69.93 (0.35)	69.94 (0.57)	69.72 (0.55)	69.75 (0.54)	69.71 (0.50)	69.71 (0.49)	69.71 (0.50)
KNN	78.67 (0.42)	78.55 (0.48)	78.45 (0.58)	78.38 (0.58)	78.42 (0.52)	78.39 (0.47)	78.43 (0.45)	78.42 (0.48)	78.51 (0.33)	78.46 (0.38)
AURC-MLP	60.24 (2.72)	59.45 (0.79)	58.67 (1.73)	57.72 (1.37)	58.78 (0.93)	57.64 (1.39)	56.77 (0.90)	56.57 (1.60)	57.39 (1.42)	56.14 (1.04)
SELE-SCR	52.93 (1.28)	52.32 (1.23)	52.34 (1.18)	52.18 (1.10)	52.10 (1.02)	52.00 (1.04)	52.01 (1.09)	51.96 (1.06)	51.98 (1.15)	51.99 (1.16)
SELE-SLR	53.05 (1.39)	52.40 (1.24)	52.39 (1.24)	52.24 (1.15)	52.12 (1.02)	52.02 (1.04)	52.03 (1.08)	51.93 (1.08)	52.01 (1.17)	51.98 (1.22)
AURC-SCR	53.15 (1.56)	52.40 (1.28)	52.41 (1.23)	52.27 (1.13)	52.16 (1.01)	52.05 (1.02)	52.06 (1.08)	52.00 (1.07)	52.05 (1.18)	51.99 (1.25)
AURC-SLR	53.11 (1.45)	52.42 (1.33)	52.41 (1.21)	52.29 (1.15)	52.15 (1.02)	52.08 (1.04)	52.06 (1.12)	51.99 (1.09)	52.06 (1.18)	**51.79** (1.34)

Qwen2.5-7B

Method	100	200	300	400	500	600	700	800	900	1000
$RL_{\max}$	34.38 (0.75)	34.38 (0.75)	34.38 (0.75)	34.38 (0.75)	34.38 (0.75)	34.38 (0.75)	34.38 (0.75)	34.38 (0.75)	34.38 (0.75)	34.38 (0.75)
$SR_{\max}$	**34.27** (0.75)	34.27 (0.75)	34.27 (0.75)	34.27 (0.75)	34.27 (0.75)	34.27 (0.75)	34.27 (0.75)	34.27 (0.75)	34.27 (0.75)	34.27 (0.75)
SR_{doctor}	38.33 (0.79)	38.33 (0.79)	38.33 (0.79)	38.33 (0.79)	38.33 (0.79)	38.33 (0.79)	38.33 (0.79)	38.33 (0.79)	38.33 (0.79)	38.33 (0.79)
SR_{ent}	38.80 (0.79)	38.80 (0.79)	38.80 (0.79)	38.80 (0.79)	38.80 (0.79)	38.80 (0.79)	38.80 (0.79)	38.80 (0.79)	38.80 (0.79)	38.80 (0.79)
$Energy$	39.35 (0.79)	39.35 (0.79)	39.35 (0.79)	39.35 (0.79)	39.35 (0.79)	39.35 (0.79)	39.35 (0.79)	39.35 (0.79)	39.35 (0.79)	39.35 (0.79)
$RL_{\text{conf-M}}$	**34.30** (0.78)	34.30 (0.78)	34.30 (0.78)	34.30 (0.78)	34.30 (0.78)	34.30 (0.78)	34.30 (0.78)	34.30 (0.78)	34.30 (0.78)	34.30 (0.78)
$RL_{\text{geo-M}}$	42.10 (1.22)	42.10 (1.22)	42.10 (1.22)	42.10 (1.22)	42.10 (1.22)	42.10 (1.22)	42.10 (1.22)	42.10 (1.22)	42.10 (1.22)	42.10 (1.22)
ViM	68.96 (0.28)	68.97 (0.39)	68.86 (0.39)	68.65 (0.39)	68.58 (0.36)	68.51 (0.33)	68.59 (0.38)	68.61 (0.31)	68.65 (0.32)	68.65 (0.31)
$SIRC$	65.04 (0.57)	64.85 (0.39)	64.72 (0.55)	64.38 (0.65)	64.09 (0.51)	63.98 (0.53)	63.97 (0.54)	63.97 (0.53)	63.94 (0.50)	63.98 (0.52)
KNN	69.07 (0.46)	69.11 (0.34)	69.14 (0.44)	68.99 (0.28)	68.97 (0.30)	68.99 (0.29)	69.06 (0.31)	69.06 (0.34)	69.10 (0.34)	69.08 (0.37)
AURC-MLP	49.06 (3.07)	47.28 (2.47)	47.60 (2.10)	45.65 (3.35)	44.73 (3.15)	43.64 (0.55)	45.21 (3.54)	44.02 (2.50)	42.79 (0.95)	44.28 (2.91)
SELE-SCR	35.36 (1.32)	34.16 (0.96)	**34.25** (1.15)	33.97 (1.28)	33.68 (0.96)	33.63 (0.93)	33.68 (0.91)	33.70 (0.96)	33.69 (0.91)	33.68 (0.88)
SELE-SLR	35.32 (1.20)	**34.02** (0.92)	**34.19** (1.09)	**33.91** (1.27)	33.67 (0.97)	33.62 (0.94)	**33.64** (0.92)	**33.64** (0.96)	**33.64** (0.91)	**33.65** (0.90)
AURC-SCR	35.61 (1.57)	34.11 (0.89)	34.26 (1.03)	33.95 (1.25)	**33.66** (0.90)	**33.60** (0.92)	**33.64** (0.92)	33.68 (0.98)	33.68 (0.94)	**33.62** (0.89)
AURC-SLR	35.93 (1.83)	**34.03** (0.83)	**34.25** (0.95)	**33.94** (1.23)	**33.63** (0.90)	**33.59** (0.91)	**33.63** (0.92)	**33.65** (1.00)	**33.66** (0.94)	**33.65** (0.90)

caveat. Source URLs: `https://github.com/Jingkang50/OpenOOD` and `https://storage.googleapis.com/openimages/web/factsfigures.html`.

CIFAR-10/100. The official University of Toronto page provides the benchmark and citation information but does not list a standalone license in its public metadata; we use the public benchmark data for evaluation and do not redistribute it. Source URL: `https://www.cs.toronto.edu/~kriz/cifar.html`.

CIFAR-10-C. Distributed through the Hendrycks robustness benchmark repository under Apache-2.0 repository terms and the Zenodo benchmark record; it is used only for evaluation. Source URLs: `https://github.com/hendrycks/robustness` and `https://zenodo.org/records/2535967`.

iWildCam. Distributed by WILDS under the Community Data License Agreement – Permissive – Version 1.0. Source URL: `https://wilds.stanford.edu/datasets/`.

PatchCamelyon. Public dataset metadata lists CC0-1.0. Source URL: `https://huggingface.co/datasets/1aurent/PatchCamelyon`.

Caltech101. Public dataset metadata from the CaltechDATA/FiftyOne listing identifies the dataset as CC BY 4.0. Source URL: `https://docs.voxel51.com/dataset_zoo/datasets/caltech101.html`.

Rendered-SST2. Torchvision/OpenAI CLIP metadata identifies the dataset source, while the Hugging Face public metadata lists the license as unknown; we use it as a benchmark split and do not redistribute the rendered images. Source URLs: `https://github.com/openai/CLIP/blob/main/data/rendered-sst2.md` and `https://huggingface.co/datasets/nateraw/rendered-sst2`.

NIH-CXR-LT. Built on NIH ChestXRay14 with labels/models from the LongTailCXR benchmark. Public metadata for the derived benchmark does not list a standalone license; we follow

Table 24: Detailed LLM calibration-size sweep on HellaSwag. **Blue**-background rows are our methods. **Green** text marks best or tied-best scores, and **Blue** text marks second-best or tied second-best scores, both at displayed precision.

Mistral-7B-v0.1

Method	100	200	300	400	500	600	700	800	900	1000
$RL_{\max}$	**37.85** (0.27)	**37.85** (0.27)	**37.85** (0.27)	**37.85** (0.27)	**37.85** (0.27)	**37.85** (0.27)	**37.85** (0.27)	**37.85** (0.27)	37.85 (0.27)	37.85 (0.27)
$SR_{\max}$	**37.83** (0.27)	**37.83** (0.27)	**37.83** (0.27)	**37.83** (0.27)	**37.83** (0.27)	**37.83** (0.27)	**37.83** (0.27)	**37.83** (0.27)	37.83 (0.27)	**37.83** (0.27)
SR_{doctor}	39.10 (0.24)	39.10 (0.24)	39.10 (0.24)	39.10 (0.24)	39.10 (0.24)	39.10 (0.24)	39.10 (0.24)	39.10 (0.24)	39.10 (0.24)	39.10 (0.24)
SR_{ent}	39.13 (0.24)	39.13 (0.24)	39.13 (0.24)	39.13 (0.24)	39.13 (0.24)	39.13 (0.24)	39.13 (0.24)	39.13 (0.24)	39.13 (0.24)	39.13 (0.24)
$Energy$	39.16 (0.23)	39.16 (0.23)	39.16 (0.23)	39.16 (0.23)	39.16 (0.23)	39.16 (0.23)	39.16 (0.23)	39.16 (0.23)	39.16 (0.23)	39.16 (0.23)
$RL_{\text{conf-M}}$	38.78 (0.35)	38.78 (0.35)	38.78 (0.35)	38.78 (0.35)	38.78 (0.35)	38.78 (0.35)	38.78 (0.35)	38.78 (0.35)	38.78 (0.35)	38.78 (0.35)
$RL_{\text{geo-M}}$	45.26 (0.37)	45.26 (0.37)	45.26 (0.37)	45.26 (0.37)	45.26 (0.37)	45.26 (0.37)	45.26 (0.37)	45.26 (0.37)	45.26 (0.37)	45.26 (0.37)
ViM	63.03 (0.92)	63.08 (1.01)	63.11 (0.97)	63.09 (0.95)	63.11 (0.96)	63.16 (1.00)	63.18 (0.99)	63.18 (0.99)	63.17 (0.97)	63.20 (0.98)
$SIRC$	57.19 (0.97)	57.32 (0.96)	57.55 (0.91)	57.53 (0.91)	57.55 (0.90)	57.65 (0.95)	57.68 (0.93)	57.67 (0.93)	57.65 (0.90)	57.69 (0.89)
KNN	63.72 (0.97)	63.66 (0.79)	63.48 (0.76)	63.49 (0.71)	63.44 (0.62)	63.48 (0.69)	63.37 (0.67)	63.33 (0.72)	63.30 (0.69)	63.34 (0.63)
AURC-MLP	46.66 (3.51)	42.07 (2.70)	41.94 (4.03)	43.67 (5.40)	40.69 (2.74)	42.54 (3.44)	42.30 (4.44)	44.89 (3.35)	44.13 (3.59)	43.47 (5.06)
SELE-SCR	38.34 (0.66)	38.07 (0.42)	37.95 (0.17)	37.95 (0.28)	37.95 (0.21)	37.93 (0.27)	37.88 (0.30)	37.87 (0.31)	37.83 (0.29)	**37.83** (0.29)
SELE-SLR	38.39 (0.71)	38.07 (0.45)	37.93 (0.18)	37.95 (0.28)	37.96 (0.21)	37.94 (0.29)	37.88 (0.32)	37.87 (0.31)	**37.80** (0.25)	37.84 (0.30)
AURC-SCR	38.33 (0.58)	38.11 (0.39)	37.93 (0.18)	37.96 (0.31)	37.97 (0.23)	37.92 (0.29)	37.88 (0.36)	37.87 (0.36)	**37.82** (0.32)	**37.82** (0.34)
AURC-SLR	38.39 (0.64)	38.12 (0.41)	37.97 (0.17)	37.97 (0.33)	37.98 (0.22)	37.94 (0.32)	37.90 (0.38)	37.87 (0.36)	37.83 (0.37)	**37.83** (0.36)

Qwen2.5-7B

Method	100	200	300	400	500	600	700	800	900	1000
$RL_{\max}$	2.48 (0.06)	2.48 (0.06)	2.48 (0.06)	2.48 (0.06)	2.48 (0.06)	2.48 (0.06)	2.48 (0.06)	2.48 (0.06)	2.48 (0.06)	2.48 (0.06)
$SR_{\max}$	2.46 (0.06)	2.46 (0.06)	2.46 (0.06)	2.46 (0.06)	2.46 (0.06)	2.46 (0.06)	2.46 (0.06)	2.46 (0.06)	2.46 (0.06)	2.46 (0.06)
SR_{doctor}	3.13 (0.05)	3.13 (0.05)	3.13 (0.05)	3.13 (0.05)	3.13 (0.05)	3.13 (0.05)	3.13 (0.05)	3.13 (0.05)	3.13 (0.05)	3.13 (0.05)
SR_{ent}	3.18 (0.05)	3.18 (0.05)	3.18 (0.05)	3.18 (0.05)	3.18 (0.05)	3.18 (0.05)	3.18 (0.05)	3.18 (0.05)	3.18 (0.05)	3.18 (0.05)
$Energy$	3.24 (0.05)	3.24 (0.05)	3.24 (0.05)	3.24 (0.05)	3.24 (0.05)	3.24 (0.05)	3.24 (0.05)	3.24 (0.05)	3.24 (0.05)	3.24 (0.05)
$RL_{\text{conf-M}}$	**2.28** (0.06)	**2.28** (0.06)	**2.28** (0.06)	**2.28** (0.06)	2.28 (0.06)	**2.28** (0.06)	**2.28** (0.06)	**2.28** (0.06)	2.28 (0.06)	2.28 (0.06)
$RL_{\text{geo-M}}$	**2.34** (0.09)	2.34 (0.09)	2.34 (0.09)	2.34 (0.09)	2.34 (0.09)	2.34 (0.09)	2.34 (0.09)	2.34 (0.09)	2.34 (0.09)	2.34 (0.09)
ViM	16.61 (1.68)	16.00 (1.37)	16.08 (1.12)	16.05 (1.18)	16.25 (1.04)	16.16 (1.13)	16.19 (1.03)	16.22 (0.91)	16.22 (0.81)	16.15 (0.84)
$SIRC$	7.42 (0.37)	7.28 (0.28)	7.23 (0.28)	7.23 (0.23)	7.22 (0.23)	7.21 (0.25)	7.18 (0.23)	7.19 (0.21)	7.17 (0.23)	7.17 (0.23)
KNN	16.50 (1.17)	16.35 (0.76)	16.53 (0.72)	16.56 (0.84)	16.73 (0.91)	16.96 (0.92)	16.97 (0.85)	16.99 (0.87)	16.96 (0.92)	16.95 (0.98)
AURC-MLP	8.00 (1.75)	7.13 (1.29)	6.50 (1.02)	6.68 (1.15)	6.24 (1.68)	5.87 (1.28)	6.09 (1.81)	5.91 (1.66)	5.47 (1.46)	5.08 (1.66)
SELE-SCR	2.46 (0.28)	2.34 (0.10)	**2.29** (0.10)	**2.29** (0.10)	2.29 (0.11)	2.29 (0.08)	**2.29** (0.08)	2.29 (0.09)	2.28 (0.07)	**2.27** (0.08)
SELE-SLR	2.48 (0.31)	2.35 (0.09)	2.30 (0.10)	2.30 (0.09)	2.29 (0.10)	2.30 (0.08)	**2.29** (0.08)	2.29 (0.08)	2.28 (0.07)	**2.27** (0.08)
AURC-SCR	2.44 (0.24)	**2.31** (0.10)	**2.28** (0.08)	**2.29** (0.09)	**2.26** (0.12)	**2.27** (0.08)	**2.28** (0.08)	**2.28** (0.11)	**2.25** (0.07)	**2.26** (0.08)
AURC-SLR	2.47 (0.28)	**2.31** (0.11)	**2.29** (0.08)	2.30 (0.10)	**2.27** (0.13)	**2.27** (0.08)	**2.28** (0.09)	**2.27** (0.10)	**2.26** (0.08)	**2.26** (0.08)

the NIH/LongTailCXR access terms and do not redistribute medical images or model weights. Source URLs: `https://github.com/VITA-Group/LongTailCXR` and `https://nihcc.app.box.com/v/ChestXray-NIHCC`.

Amazon Reviews. Based on the public Amazon review data used in [37]. Public metadata for the downstream Amazon review assets we checked lists the license as unknown; we cite the source and do not redistribute review text. Source URLs: `https://nijianmo.github.io/amazon/index.html` and `https://aclanthology.org/D19-1018/`.

EyePACS. Access is governed by Kaggle/EyePACS competition or provider terms; redistribution may be restricted, so we do not redistribute images. Source URL: `https://www.kaggle.com/c/diabetic-retinopathy-detection`.

ADE-Corpus-V2. Hugging Face dataset metadata lists the license as unknown; we cite the original ADE-Corpus-V2 source and do not redistribute the corpus text. Source URL: `https://huggingface.co/datasets/ade-benchmark-corpus/ade_corpus_v2`.

MMLU. Public Hugging Face metadata lists MIT. Source URL: `https://huggingface.co/datasets/cais/mmlu`.

MMLU-Pro. Public Hugging Face metadata lists MIT. Source URL: `https://huggingface.co/datasets/TIGER-Lab/MMLU-Pro`.

HellaSwag. Public metadata and the original repository list MIT. Source URL: `https://huggingface.co/datasets/Rowan/hellaswag`.

ARC-Easy. The AI2 ARC dataset metadata lists CC BY-SA 4.0. Source URL: `https://huggingface.co/datasets/allenai/ai2_arc`.

Pretrained/fixed models and code assets.

Table 25: Detailed LLM calibration-size sweep on ARC-Easy. **Blue**-background rows are our methods. **Green** text marks best or tied-best scores, and **Blue** text marks second-best or tied second-best scores, both at displayed precision.

Mistral-7B-v0.1

Method	100	200	300	400	500	600	700	800	900	1000
$RL_{\max}$	2.66 (0.17)	2.66 (0.17)	2.66 (0.17)	2.66 (0.17)	2.66 (0.17)	2.66 (0.17)	2.66 (0.17)	2.66 (0.17)	2.66 (0.17)	2.66 (0.17)
$SR_{\max}$	2.63 (0.16)	2.63 (0.16)	2.63 (0.16)	2.63 (0.16)	2.63 (0.16)	2.63 (0.16)	2.63 (0.16)	2.63 (0.16)	2.63 (0.16)	2.63 (0.16)
SR_{doctor}	3.12 (0.25)	3.12 (0.25)	3.12 (0.25)	3.12 (0.25)	3.12 (0.25)	3.12 (0.25)	3.12 (0.25)	3.12 (0.25)	3.12 (0.25)	3.12 (0.25)
SR_{ent}	3.19 (0.26)	3.19 (0.26)	3.19 (0.26)	3.19 (0.26)	3.19 (0.26)	3.19 (0.26)	3.19 (0.26)	3.19 (0.26)	3.19 (0.26)	3.19 (0.26)
$Energy$	3.27 (0.29)	3.27 (0.29)	3.27 (0.29)	3.27 (0.29)	3.27 (0.29)	3.27 (0.29)	3.27 (0.29)	3.27 (0.29)	3.27 (0.29)	3.27 (0.29)
$RL_{\text{conf-M}}$	**2.48** (0.15)	**2.48** (0.15)	**2.48** (0.15)	**2.48** (0.15)	**2.48** (0.15)	2.48 (0.15)	2.48 (0.15)	2.48 (0.15)	2.48 (0.15)	2.48 (0.15)
$RL_{\text{geo-M}}$	**2.50** (0.11)	**2.50** (0.11)	**2.50** (0.11)	2.50 (0.11)	2.50 (0.11)	2.50 (0.11)	2.50 (0.11)	2.50 (0.11)	2.50 (0.11)	2.50 (0.11)
ViM	12.42 (0.61)	12.71 (0.77)	13.12 (1.00)	13.21 (0.93)	13.16 (0.93)	13.08 (1.18)	13.05 (1.24)	13.08 (1.24)	13.04 (1.19)	13.11 (1.29)
$SIRC$	4.88 (0.33)	4.84 (0.36)	4.93 (0.35)	4.93 (0.35)	4.91 (0.35)	4.85 (0.38)	4.83 (0.39)	4.84 (0.38)	4.84 (0.36)	4.86 (0.35)
KNN	13.84 (0.89)	13.61 (1.36)	13.45 (1.36)	13.63 (1.32)	13.62 (1.40)	13.09 (1.48)	13.02 (1.46)	13.25 (1.41)	13.20 (1.47)	13.12 (1.54)
AURC-MLP	7.48 (0.86)	6.27 (1.43)	5.63 (1.33)	6.57 (1.96)	5.47 (1.23)	5.06 (1.62)	5.66 (1.77)	6.21 (1.96)	6.76 (2.15)	5.95 (2.08)
SELE-SCR	2.65 (0.21)	2.53 (0.11)	2.52 (0.13)	**2.48** (0.09)	**2.48** (0.09)	**2.45** (0.12)	**2.44** (0.13)	**2.44** (0.13)	**2.44** (0.12)	**2.44** (0.13)
SELE-SLR	2.66 (0.24)	2.55 (0.12)	2.54 (0.13)	**2.49** (0.10)	**2.49** (0.09)	**2.46** (0.12)	**2.44** (0.13)	**2.44** (0.12)	**2.45** (0.13)	**2.44** (0.13)
AURC-SCR	2.73 (0.18)	2.56 (0.09)	**2.50** (0.09)	**2.48** (0.09)	**2.48** (0.10)	**2.45** (0.13)	**2.45** (0.14)	2.47 (0.14)	2.46 (0.14)	**2.45** (0.14)
AURC-SLR	2.74 (0.24)	2.57 (0.10)	2.51 (0.07)	2.51 (0.10)	**2.49** (0.10)	**2.46** (0.13)	2.46 (0.14)	**2.46** (0.13)	**2.45** (0.13)	**2.45** (0.15)

Qwen2.5-7B

Method	100	200	300	400	500	600	700	800	900	1000
$RL_{\max}$	0.37 (0.12)	0.37 (0.12)	0.37 (0.12)	0.37 (0.12)	0.37 (0.12)	0.37 (0.12)	0.37 (0.12)	0.37 (0.12)	0.37 (0.12)	0.37 (0.12)
$SR_{\max}$	0.37 (0.11)	0.37 (0.11)	0.37 (0.11)	0.37 (0.11)	0.37 (0.11)	0.37 (0.11)	0.37 (0.11)	0.37 (0.11)	0.37 (0.11)	0.37 (0.11)
SR_{doctor}	0.57 (0.13)	0.57 (0.13)	0.57 (0.13)	0.57 (0.13)	0.57 (0.13)	0.57 (0.13)	0.57 (0.13)	0.57 (0.13)	0.57 (0.13)	0.57 (0.13)
SR_{ent}	0.60 (0.13)	0.60 (0.13)	0.60 (0.13)	0.60 (0.13)	0.60 (0.13)	0.60 (0.13)	0.60 (0.13)	0.60 (0.13)	0.60 (0.13)	0.60 (0.13)
$Energy$	0.62 (0.13)	0.62 (0.13)	0.62 (0.13)	0.62 (0.13)	0.62 (0.13)	0.62 (0.13)	0.62 (0.13)	0.62 (0.13)	0.62 (0.13)	0.62 (0.13)
$RL_{\text{conf-M}}$	**0.32** (0.13)	**0.32** (0.13)	**0.32** (0.13)	**0.32** (0.13)	**0.32** (0.13)	**0.32** (0.13)	**0.32** (0.13)	**0.32** (0.13)	**0.32** (0.13)	**0.32** (0.13)
$RL_{\text{geo-M}}$	**0.33** (0.11)	**0.33** (0.11)	**0.33** (0.11)	**0.33** (0.11)	**0.33** (0.11)	**0.33** (0.11)	**0.33** (0.11)	**0.33** (0.11)	**0.33** (0.11)	**0.33** (0.11)
ViM	3.67 (0.89)	3.32 (0.54)	3.19 (0.36)	3.11 (0.31)	3.04 (0.35)	2.95 (0.34)	2.96 (0.31)	3.00 (0.31)	3.05 (0.27)	3.10 (0.28)
$SIRC$	1.05 (0.15)	1.02 (0.19)	1.01 (0.17)	1.00 (0.16)	0.97 (0.15)	0.96 (0.14)	0.95 (0.14)	0.95 (0.13)	0.97 (0.14)	0.98 (0.14)
KNN	1.91 (0.35)	2.06 (0.44)	2.04 (0.29)	2.08 (0.33)	2.16 (0.29)	2.11 (0.24)	2.12 (0.25)	2.06 (0.22)	2.06 (0.22)	2.04 (0.26)
AURC-MLP	2.48 (0.28)	2.24 (0.83)	1.99 (0.69)	1.95 (0.71)	2.21 (0.70)	2.17 (0.77)	1.89 (0.72)	1.64 (0.73)	1.65 (0.64)	1.83 (0.55)
SELE-SCR	0.38 (0.16)	**0.33** (0.11)	**0.33** (0.11)	0.34 (0.12)	0.34 (0.12)	0.34 (0.12)	**0.33** (0.12)	**0.33** (0.12)	**0.33** (0.12)	**0.32** (0.12)
SELE-SLR	0.41 (0.15)	0.34 (0.11)	**0.33** (0.11)	0.34 (0.12)	0.34 (0.12)	**0.33** (0.12)	**0.33** (0.12)	**0.33** (0.12)	**0.33** (0.12)	**0.32** (0.11)
AURC-SCR	0.38 (0.14)	**0.33** (0.11)	0.34 (0.11)	0.34 (0.11)	0.34 (0.11)	**0.33** (0.11)	**0.32** (0.11)	**0.33** (0.11)	**0.32** (0.11)	**0.32** (0.11)
AURC-SLR	0.38 (0.14)	**0.33** (0.11)	**0.33** (0.11)	0.34 (0.12)	0.34 (0.12)	**0.33** (0.12)	**0.33** (0.12)	**0.32** (0.12)	**0.32** (0.11)	**0.32** (0.12)

Table 26: Ablation study results ($\alpha = 1.0$) across datasets with different numbers of base selectors at calibration size 1000. Entries are mean (standard deviation) over 5 repeats. **Blue**-background rows are our methods. **Green** numbers indicate the best mean score per (dataset, #selectors); **Blue** numbers indicate the second-best mean. Abbreviations follow Tab. 2.

Method	ADE	AMZ	CAL	C10	C1C	C100	EYE	IMN	INO	WID	WOD	OIO	NIH	PCM	SST	#Selectors
SELE-SCR	55.50 (1.34)	12.15 (0.24)	0.75 (0.13)	0.98 (0.12)	6.54 (0.07)	24.07 (0.26)	9.20 (0.22)	**3.22** (0.10)	4.59 (0.08)	9.43 (0.38)	8.57 (0.06)	11.28 (0.14)	72.47 (0.99)	8.43 (0.25)	9.11 (0.24)	10
SELE-SCR	55.17 (1.22)	12.18 (0.27)	0.84 (0.16)	0.98 (0.12)	6.60 (0.18)	24.09 (0.29)	9.20 (0.22)	**3.22** (0.10)	4.59 (0.08)	9.42 (0.38)	8.62 (0.10)	11.26 (0.15)	72.89 (1.03)	8.43 (0.26)	9.11 (0.24)	7
SELE-SCR	55.86 (0.75)	12.22 (0.23)	0.85 (0.21)	0.98 (0.10)	6.54 (0.08)	24.01 (0.21)	9.64 (0.15)	**3.22** (0.10)	4.58 (0.08)	9.59 (0.45)	8.66 (0.04)	11.26 (0.15)	72.88 (1.03)	9.32 (0.53)	9.10 (0.27)	4
SELE-SLR	**53.16** (1.23)	**12.13** (0.25)	**0.72** (0.13)	0.98 (0.11)	**6.53** (0.08)	23.91 (0.31)	**9.19** (0.20)	**3.22** (0.09)	**4.58** (0.06)	9.40 (0.41)	8.49 (0.17)	**11.20** (0.11)	**72.39** (1.04)	**6.46** (0.16)	9.10 (0.24)	10
SELE-SLR	55.42 (1.25)	**12.15** (0.24)	0.79 (0.19)	0.98 (0.10)	**6.53** (0.08)	23.93 (0.29)	9.20 (0.21)	**3.21** (0.09)	**4.58** (0.05)	9.39 (0.40)	8.45 (0.16)	**11.13** (0.13)	72.89 (1.07)	**6.46** (0.17)	9.11 (0.24)	7
SELE-SLR	**55.83** (0.73)	12.22 (0.23)	0.83 (0.24)	0.97 (0.08)	**6.52** (0.07)	23.89 (0.29)	**9.57** (0.13)	**3.21** (0.10)	**4.58** (0.08)	9.59 (0.50)	**8.51** (0.22)	**11.25** (0.13)	72.86 (1.04)	9.33 (0.53)	**9.07** (0.26)	4
AURC-SCR	55.25 (1.67)	12.14 (0.25)	0.73 (0.11)	**0.97** (0.07)	6.54 (0.06)	23.79 (0.24)	9.21 (0.16)	3.24 (0.09)	4.58 (0.08)	**9.32** (0.34)	**8.36** (0.09)	11.23 (0.13)	72.45 (0.93)	8.33 (0.19)	**9.08** (0.18)	10
AURC-SCR	**55.04** (1.51)	12.15 (0.24)	0.86 (0.22)	**0.97** (0.08)	6.53 (0.07)	**23.80** (0.25)	9.23 (0.20)	3.24 (0.09)	4.58 (0.08)	9.31 (0.30)	**8.34** (0.08)	11.23 (0.12)	72.86 (1.03)	8.32 (0.18)	**9.05** (0.18)	7
AURC-SCR	**55.83** (0.73)	**12.21** (0.23)	0.84 (0.22)	0.97 (0.07)	6.55 (0.05)	**23.79** (0.24)	9.66 (0.12)	3.22 (0.10)	4.58 (0.09)	9.56 (0.40)	8.58 (0.08)	**11.23** (0.13)	72.85 (1.02)	**9.13** (0.25)	9.08 (0.16)	4
AURC-SLR	**53.21** (1.81)	**12.12** (0.26)	**0.68** (0.13)	**0.97** (0.07)	**6.48** (0.06)	**23.76** (0.26)	9.23 (0.21)	3.25 (0.09)	**4.56** (0.10)	**9.07** (0.46)	**8.16** (0.08)	**11.07** (0.10)	**72.25** (0.98)	**6.45** (0.22)	**9.07** (0.17)	10
AURC-SLR	55.07 (1.55)	**12.15** (0.25)	**0.67** (0.12)	**0.97** (0.07)	**6.49** (0.04)	**23.77** (0.27)	9.23 (0.23)	3.24 (0.09)	**4.57** (0.09)	**9.10** (0.36)	**8.21** (0.13)	**11.09** (0.12)	**72.74** (1.00)	**6.43** (0.21)	**9.06** (0.16)	7
AURC-SLR	55.84 (0.73)	**12.21** (0.23)	**0.77** (0.15)	**0.96** (0.07)	**6.49** (0.05)	**23.77** (0.26)	9.58 (0.20)	3.22 (0.10)	4.58 (0.08)	**9.31** (0.32)	**8.28** (0.06)	11.28 (0.11)	**72.71** (0.97)	**9.01** (0.32)	**9.07** (0.15)	4

Table 27: Compute-resource summary for reproducing the experimental pipeline. Runtime estimates are per submitted job or configuration.

Stage	Worker type	Memory	Runtime
Image/NLP feature, logit, and base-score collection	1 NVIDIA A100 GPU (40 GB) and 4 CPU cores	64 GB RAM	At most 3 hours per dataset/model pass
LLM feature/logit collection	1 NVIDIA A100 GPU (40 GB) and 4 CPU cores	64 GB RAM	At most 3 hours per benchmark/model pass for a 7B model in fp16
Ensemble training and risk–coverage evaluation	CPU-only, 4 CPU cores	16 GB RAM	At most 30 minutes per $\langle$dataset, calibration size, β, seed, method$\rangle$ configuration
Batch ensemble sweeps	CPU-only, 8 CPU cores with 8 parallel workers	16 GB RAM	At most 2 hours per batch job

timm/ResNet. The timm package is Apache-2.0; pretrained weights are used through their source package terms. Source URL: `https://github.com/huggingface/pytorch-image-models`.

FLYP. The official FLYP repository is MIT-licensed. Source URL: `https://github.com/locuslab/FLYP`.

LONGTAILCXR. The LongTailCXR repository provides code, labels, and model links but does not expose a standalone license in the public repository metadata we checked; we cite the benchmark and do not redistribute weights or medical data. Source URL: `https://github.com/VITA-Group/LongTailCXR`.

LISA. LISA refers to the selective-augmentation training method of Yao et al. [48]. For the Amazon benchmark, we follow the generalized selective-classification protocol of Liang et al. [4]. The official LISA source code is available at `https://github.com/huaxiuyao/LISA`; we do not redistribute model weights.

SpanBERT. The original SpanBERT repository states that SpanBERT is CC BY-NC 4.0 and that this license applies to the pretrained models. The ADE-finetuned Hugging Face model used in this paper does not list a separate license in its public metadata, so we treat it as governed by the SpanBERT/model-provider terms and do not redistribute it. Source URLs: `https://github.com/facebookresearch/SpanBERT` and `https://huggingface.co/abhibisht89/spanbert-large-cased-finetuned-ade_corpus_v2`.

Mistral-7B-v0.1. The Hugging Face model card lists Apache-2.0. Source URL: `https://huggingface.co/mistralai/Mistral-7B-v0.1`.

Qwen2.5-7B. The Hugging Face model card lists Apache-2.0. Source URL: `https://huggingface.co/Qwen/Qwen2.5-7B`.

Baseline confidence scores and released code. The post-hoc confidence scores in App. C are implemented from published definitions and cited sources, including Energy [16], KNN [28], ViM [19], SIRC [24], and the benchmark protocol in [4]. The supplementary reproduction code is a new asset released by the authors and will include its own license file and third-party attribution notes.

F Broader Impacts

Better post-hoc selective classification can help practitioners abstain on low-confidence predictions from fixed classifiers, which may improve reliability in settings where retraining is impractical. This can be useful in high-stakes deployment contexts such as biomedical decision support, where a model may flag uncertain diagnostic predictions for clinician review, or financial decision support, where uncertain risk or fraud predictions may require additional human verification before action is taken.

At the same time, a learned confidence ranking should not be interpreted as a guarantee of correctness. It may inherit biases or systematic blind spots from the base classifier and can become miscalibrated under deployment shift. In high-stakes settings such as healthcare, finance, or other decision pipelines that affect individual access to services, inappropriate reliance on selective predictions could still

produce unfair or harmful outcomes. Such uses require domain-specific validation, monitoring, human oversight, and revalidation when the deployment distribution changes.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and Sec. 1 state the paper's main contributions and scope, including the post-hoc setting, the SoftRank-AURC objective, theoretical guarantee, and empirical evaluation. The limitations are discussed separately in Sec. 6.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper includes a dedicated limitations discussion in Sec. 6, covering the fixed-classifier setting, reliance on deployment calibration data, and the quadratic cost of pairwise SoftRank-AURC comparisons.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: The theoretical assumptions and main oracle inequality are stated in Sec. 3.3. Complete proofs and supporting lemmas are provided in App. B.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The method, training objectives, normalization procedure, calibration protocol, datasets, models, and hyperparameter choices are described in Secs. 3 and 4 and Apps. A.1, A.3 and C. Additional calibration-size and selector-count results are reported in the appendix.

Guidelines:

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

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: We provide anonymized source code and reproduction instructions in the supplementary material. The experiments use existing public or benchmark datasets and fixed pretrained models; dataset/model sources, calibration/test splits, preprocessing, hyperparameters, and evaluation protocols are described in Sec. 4 and App. C.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The experimental protocol, data splits, validation procedure, optimizer, and tuning grids are described in Sec. 4. Dataset/model details and extended result tables are provided in App. C.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The main experimental tables report point estimates on the specified calibration/validation/test splits. The LLM calibration-size tables report means over 5 random seeds. The corresponding standard deviation results are reported in the App. C

   Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We disclose the compute used for each stage of the pipeline in App. D. The image/NLP and LLM feature/logit collection stages use one NVIDIA A100 GPU per job, while the main ensemble training and risk–coverage evaluation are CPU-only after logits, features, and base confidence scores are cached. We also report per-job runtimes, total final-run compute, and the approximate additional compute used for preliminary and abandoned runs.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics and believe the research conforms to it. The paper uses existing public or benchmark datasets and fixed pretrained models for post-hoc selective-classification evaluation.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The paper studies post-hoc selective classification for improving reliability of fixed classifiers. Potential deployment benefits and risks, including calibration dependence, distribution shift, high-stakes deployment examples, and the inability to correct systematic base-model errors, are discussed in Secs. 1 and 6 and App. F.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not introduce or release new high-risk datasets, pretrained language models, image generators, or other models with substantial misuse risk. The LLM experiments evaluate existing open-weight models in a multiple-choice classification setting.

    Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: Existing datasets, pretrained/fixed models, and baseline confidence scores are credited through citations and source entries in Sec. 4 and App. C. License, terms-of-use, and source-URL notes for third-party assets, including assets whose public metadata does not list a standalone license, are summarized in App. E. The supplementary code is a new asset and will be released with its own license and third-party attribution notes.

Guidelines:

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

13. **New assets**

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

Answer: [Yes]

Justification: The new asset released with the submission is anonymized source code, not a new dataset or pretrained model. The supplementary material documents the code with usage instructions, dependencies, and license information.

Guidelines:

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

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

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

Answer: [N/A]

Justification: The paper does not conduct new crowdsourcing experiments or new research with human subjects. The empirical evaluation uses existing public or benchmark datasets and fixed pretrained models.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [N/A]

    Justification: The paper does not involve new crowdsourcing experiments or new human-subjects research, so IRB approval or equivalent review is not applicable.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: The paper describes the LLM multiple-choice experiments in Sec. 4.2 and provides additional benchmark statistics and calibration-size results in App. C. LLMs are used as fixed classifiers for evaluation, not as part of the proposed method development.

    Guidelines:

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