Learning to Defer to Multiple Experts: Consistent Surrogate Losses, Confidence Calibration, and Conformal Ensembles

Rajeev Verma* University of Amsterdam
Daniel Barrejón* Universidad Carlos III de Madrid
Eric Nalisnick University of Amsterdam
*Equal contribution, order determined by coin flip

Abstract

We study the statistical properties of *learning to defer* (L2D) to multiple experts. In particular, we address the open problems of deriving a consistent surrogate loss, confidence calibration, and principled ensembling of experts. Firstly, we derive two consistent surrogates—one based on a softmax parameterization, the other on a one-vs-all (OvA) parameterization—that are analogous to the single expert losses proposed by Mozannar and Sontag (2020) and Verma and Nalisnick (2022), respectively. We then study the frameworks' ability to estimate $\mathbb{P}(\mathrm{m}_j = \mathrm{y}|\boldsymbol{x})$, the probability that the jth expert will correctly predict the label for $\boldsymbol{x}$. Theory shows the softmax-based loss causes mis-calibration to propagate between the estimates while the OvA-based loss does not (though in practice, we find there are trade offs). Lastly, we propose a conformal inference technique that chooses a subset of experts to query when the system defers. We perform empirical validation on tasks for galaxy, skin lesion, and hate speech classification.

1 INTRODUCTION

Solving complex problems often requires the involvement of multiple experts (Fay et al., 2006). These experts may have non-overlapping specialties, such as in a large construction project that requires the advice of engineers, architects, geologists, lawyers, etc. Or perhaps the difficulty of the task requires multiple opinions, as when a team of doctors consults on a difficult medical diagnosis. Thus, modern *hybrid intelligent* (HI) systems (Kamar, 2016; Dellermann et al., 2019; Akata et al., 2020)—so called because they combine computational and human decision making—need to support the participation of multiple experts.

Learning to defer (L2D) (Madras et al., 2018) presents an elegant framework for implementing HI systems. A *rejector* model acts as a meta-classifier, predicting whether the downstream classifier or human is more likely to make the correct decision for a given input. Yet existing L2D frameworks do not obviously accommodate additional experts. For instance, the rejector's job becomes more challenging when there are multiple experts. It has two decisions to make: *when* to defer and *to which* expert. The latter decision is equally important, and thus verifying that the rejector can accurately monitor expert quality is essential for safe and effective deployment.

In this paper, we develop the statistical foundations of multiclass L2D with multiple experts. Specifically, we address the following open problems: deriving a consistent surrogate loss, studying whether these systems are confidence calibrated, and developing a principled technique for ensembling expert decisions. The first and second contributions ensure the soundness of the optimization problem and resulting downstream decision making. Our third contribution, expert ensembles, follows from our study of calibration, as we propose a conformal inference procedure for selecting a subset of experts. We empirically validate our methods on the tasks of image classification, galaxy categorization, skin lesion diagnosis, and hate speech detection. We find that our consistent losses result in superior accuracy and calibration when compared to existing systems based on (inconsistent) mixtures of experts (Hemmer et al., 2022).

2 LEARNING TO DEFER

Mozannar and Sontag (2020) and Verma and Nalisnick (2022) proposed the only known consistent (surrogate) loss functions for multiclass L2D. Thus we focus on their formulations. We provide the technical background of L2D in this section before moving on to the multi-expert setting.

Data We first define the data for multiclass L2D with one expert. Let $\mathcal{X}$ denote the feature space, and let $\mathcal{Y}$ denote the

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

label space, which we assume to be a categorical encoding of multiple (K) classes. $\mathbf{x}_n \in \mathcal{X}$ denotes a feature vector, and $\mathrm{y}_n \in \mathcal{Y}$ denotes the associated class defined by $\mathcal{Y}$ (1 of K). The L2D problem also assumes that we have access to (human) expert demonstrations. Denote the expert's prediction space as $\mathcal{M}$, which is usually taken to be equal to the label space: $\mathcal{M} = \mathcal{Y}$. Expert demonstrations are denoted $\mathrm{m}_n \in \mathcal{M}$ for the associated features $\mathbf{x}_n$. The combined N-element training sample is $\mathcal{D} = \{\boldsymbol{x}_n, y_n, m_n\}_{n=1}^N$.

Models and Learning Mozannar and Sontag (2020)'s and Verma and Nalisnick (2022)'s L2D frameworks compose two models: a classifier and a rejector (Cortes et al., 2016a,b). Denote the *classifier* as $h : \mathcal{X} \to \mathcal{Y}$ and the *rejector* as $r : \mathcal{X} \to \{0, 1\}$. When $r(\mathbf{x}) = 0$, the classifier makes the decision. When $r(\mathbf{x}) = 1$, the system defers the decision to the human. When the classifier makes the prediction, then the system incurs a loss $\ell(h(\boldsymbol{x}), y)$. When the human makes the prediction (i.e. $r(\boldsymbol{x}) = 1$), the system incurs a loss $\ell_{\text{exp}}(m, y)$. Using the rejector to combine these losses, we have the overall classifier-rejector loss:

$$L(h, r) = \mathbb{E}_{\mathbf{x},\mathrm{y},\mathrm{m}} \left[(1 - r(\mathbf{x}))\, \ell(h(\mathbf{x}), \mathrm{y}) \;+\; r(\mathbf{x})\, \ell_{\text{exp}}(\mathrm{m}, \mathrm{y})\right] \tag{1}$$

where the rejector is acting as an indicator function that controls which loss to use. While this formulation is valid for general losses, the canonical $0-1$ loss is of special interest for classification tasks:

$$L_{0-1}(h, r) = \mathbb{E}_{\mathbf{x},\mathrm{y},\mathrm{m}} \left[(1 - r(\mathbf{x}))\, \mathbb{I}[h(\mathbf{x}) \neq \mathrm{y}] \;+\; r(\mathbf{x})\, \mathbb{I}[\mathrm{m} \neq \mathrm{y}]\right] \tag{2}$$

where $\mathbb{I}$ denotes an indicator function that checks if the prediction and label are equal. Upon minimization, the resulting Bayes optimal classifier and rejector satisfy:

$$\begin{aligned} h^*(\boldsymbol{x}) &= \arg\max_{y \in \mathcal{Y}} \; \mathbb{P}(\mathrm{y} = y | \boldsymbol{x}), \\ r^*(\boldsymbol{x}) &= \mathbb{I}\left[\mathbb{P}(\mathrm{m} = \mathrm{y} | \boldsymbol{x}) \geq \max_{y \in \mathcal{Y}} \mathbb{P}(\mathrm{y} = y | \boldsymbol{x})\right], \end{aligned} \tag{3}$$

where $\mathbb{P}(\mathrm{y}|\boldsymbol{x})$ is the probability of the label under the data generating process, and $\mathbb{P}(\mathrm{m} = \mathrm{y}|\boldsymbol{x})$ is the probability that the expert is correct. The expert likely will have additional knowledge not available to the classifier, which possibly allows the expert to outperform the Bayes optimal classifier.

Softmax Surrogate Mozannar and Sontag (2020) proposed the first consistent surrogate loss for L_{0-1}, meaning that its minimizers agree with the Bayes optimal solutions in Equation 3. They accomplish this by first unifying the classifier and rejector via an augmented label space that includes the rejection option. Formally, this label space is defined as $\mathcal{Y}^\perp = \mathcal{Y} \cup \{\perp\}$ where $\perp$ denotes the rejection option. Secondly, Mozannar and Sontag (2020) use a reduction to cost sensitive learning that ultimately resembles the cross-entropy loss for a softmax parameterization. Let $g_k : \mathcal{X} \mapsto \mathbb{R}$ for $k \in [1, K]$ where k denotes the class index, and let $g_\perp : \mathcal{X} \mapsto \mathbb{R}$ denote the rejection ($\perp$) option. These $K + 1$ functions are then combined in the following softmax-based, point-wise surrogate loss:

$$\begin{aligned} \phi_{\text{SM}}(g_1, \ldots, g_K, g_\perp; \boldsymbol{x}, y, m) = \\ - \log\left(\frac{\exp\{g_y(\boldsymbol{x})\}}{\sum_{y' \in \mathcal{Y}^\perp} \exp\{g_{y'}(\boldsymbol{x})\}}\right) \\ - \mathbb{I}[m = y]\; \log\left(\frac{\exp\{g_\perp(\boldsymbol{x})\}}{\sum_{y' \in \mathcal{Y}^\perp} \exp\{g_{y'}(\boldsymbol{x})\}}\right). \end{aligned} \tag{4}$$

The intuition is that the first term maximizes the function g_k associated with the true label. The second term then maximizes the rejection function $g_\perp$ but only if the expert's prediction is correct. At test time, the classifier is obtained by taking the maximum over $k \in [1, K]$: $\hat{y} = h(\boldsymbol{x}) = \arg\max_{k \in [1,K]} g_k(\boldsymbol{x})$. The rejection function is similarly formulated as $r(\boldsymbol{x}) = \mathbb{I}[g_\perp(\boldsymbol{x}) \geq \max_k g_k(\boldsymbol{x})]$.

One-vs-All Surrogate Verma and Nalisnick (2022) proposed an alternative consistent surrogate for multiclass L2D based on a one-vs-all (OvA) formulation. Again assume we have $K + 1$ functions $g_1(\mathbf{x}), \ldots, g_K(\mathbf{x}), g_\perp(\mathbf{x})$ such that $g : \mathcal{X} \mapsto \mathbb{R}$. Their one-vs-all (OvA) surrogate loss has the point-wise form:

$$\begin{aligned} \psi_{\text{OvA}}(g_1, \ldots, g_K, g_\perp; \boldsymbol{x}, y, m) = \\ \phi[g_y(\boldsymbol{x})] + \sum_{y' \in \mathcal{Y}, y' \neq y} \phi[-g_{y'}(\boldsymbol{x})] \;+ \\ \phi[-g_\perp(\boldsymbol{x})] + \mathbb{I}[m = y]\,(\phi[g_\perp(\boldsymbol{x})] - \phi[-g_\perp(\boldsymbol{x})]) \end{aligned} \tag{5}$$

where $\phi : \{\pm 1\} \times \mathbb{R} \mapsto \mathbb{R}_+$ is a binary surrogate loss. For instance, when ϕ is the logistic loss, we have $\phi[f(\boldsymbol{x})] = \log(1 + \exp\{-f(\boldsymbol{x})\})$. The classifier and rejector are computed exactly the same as for the softmax loss. The motivation for this loss is that it produces better calibrated systems than those produced by the softmax-based loss. The softmax loss has a degenerate parameterization that causes it, in practice, to overestimate the expert's probability of correctness (Verma and Nalisnick, 2022).

3 L2D TO MULTIPLE EXPERTS

We now turn to the multi-expert setting, deriving two consistent surrogate losses that are analogous to Mozannar and Sontag (2020)'s and Verma and Nalisnick (2022)'s single-expert loss functions.

Data Again let $\mathbf{x}_n \in \mathcal{X}$ and $\mathrm{y}_n \in \mathcal{Y}$ be the feature and label respectively, as defined above. Now let there be J experts, and denote each expert's prediction space as $\mathcal{M}_j$ (which again we will assume is equal to the label space: $\mathcal{M}_j = \mathcal{Y} \;\; \forall j$). The expert demonstrations

are denoted $\mathrm{m}_{n,j} \in \mathcal{M}_j$ for the associated features $\mathbf{x}_n$. The combined N-element training sample is then denoted $\mathcal{D} = \{\boldsymbol{x}_n, y_n, m_{n,1}, \ldots, m_{n,J}\}_{n=1}^N$.

Models Again we use the classifier-rejector formulation. The classifier (h) is unchanged from the single-expert setting. The rejector, on the other hand, must be modified. In L2D with one expert, the rejector makes a binary decision—to defer or not. In multi-expert L2D, the rejector also must choose *to which* expert to assign the instance. Hence let the rejector be denoted $r : \mathcal{X} \to \{0, 1, \ldots, J\}$. When $r(\mathbf{x}) = 0$, the classifier makes the decision. When $r(\mathbf{x}) = j$, the system defers the decision to the jth expert.

Learning Again the learning objective is the $0-1$ loss. We can re-write Equation 2 for the multi-expert setting as:

$$L_{0-1}(h, r) = \mathbb{E}_{\mathbf{x},\mathrm{y},\{\mathrm{m}_j\}_{j=1}^J}\Bigg[\mathbb{I}[r(\mathbf{x}) = 0]\,\mathbb{I}[h(\mathbf{x}) \neq \mathrm{y}] + \sum_{j=1}^{J} \mathbb{I}[r(\mathbf{x}) = j]\,\mathbb{I}[\mathrm{m}_j \neq \mathrm{y}]\Bigg] \tag{6}$$

The corresponding Bayes optimal classifier and rejector are:

$$h^*(\boldsymbol{x}) = \arg\max_{y \in \mathcal{Y}} \mathbb{P}(\mathrm{y} = y|\boldsymbol{x}),$$
$$r^*(\boldsymbol{x}) = \begin{cases} 0 \text{ if } \mathbb{P}(\mathrm{y} = h^*(\boldsymbol{x})|\boldsymbol{x}) > \mathbb{P}(\mathrm{m}_{j'} = \mathrm{y}|\boldsymbol{x}) \;\; \forall j' \\ \arg\max_{j \in [1,J]} \mathbb{P}(\mathrm{m}_j = \mathrm{y}|\boldsymbol{x}) \quad \text{otherwise,} \end{cases} \tag{7}$$

where $\mathbb{P}(\mathrm{y}|\boldsymbol{x})$ is again the probability of the label under the data generating process and $\mathbb{P}(\mathrm{m}_j = \mathrm{y}|\boldsymbol{x})$ is the true probability that the jth expert is correct. We provide the derivation of this rule in Section A.1.

3.1 Softmax Surrogate Loss

Given the preceding definitions, we can now define the multi-expert analog of the softmax-based surrogate loss. Define the augmented label space as $\mathcal{Y}^\perp = \mathcal{Y} \cup \{\perp_1, \ldots, \perp_J\}$ where $\perp_j$ denotes the decision to defer to the jth expert. Let the classifier be composed of K functions: $g_k : \mathcal{X} \mapsto \mathbb{R}$ for $k \in [1, K]$ where k denotes the class index. Then let the rejector be implemented with J functions: $g_{\perp,j} : \mathcal{X} \mapsto \mathbb{R}$ for $j \in [1, J]$ where j is the expert index. We propose to combine these $K + J$ functions via the following softmax-parameterized surrogate loss:

$$\Phi_{\mathrm{SM}}^J(g_1, \ldots, g_K, g_{\perp,1}, \ldots, g_{\perp,J}; \boldsymbol{x}, y, m_1, \ldots, m_J) = -\log\left(\frac{\exp\{g_y(\boldsymbol{x})\}}{\sum_{y' \in \mathcal{Y}^\perp} \exp\{g_{y'}(\boldsymbol{x})\}}\right) - \sum_{j=1}^{J} \mathbb{I}[m_j = y]\,\log\left(\frac{\exp\{g_{\perp,j}(\boldsymbol{x})\}}{\sum_{y' \in \mathcal{Y}^\perp} \exp\{g_{y'}(\boldsymbol{x})\}}\right). \tag{8}$$

The intuition is that the first term maximizes the function g_k associated with the true label. The second term maximizes the rejection function $g_{\perp,j}$ but only if the jth expert's prediction is correct. At test time, the classifier is obtained by taking the maximum over $k \in [1, K]$: $\hat{y} = h(\boldsymbol{x}) = \arg\max_{k \in [1,K]} g_k(\boldsymbol{x})$. The rejection function is similarly formulated as

$$r(\boldsymbol{x}) = \begin{cases} 0 \;\; \text{if} \;\; g_{h(\boldsymbol{x})} > g_{\perp,j'} \;\; \forall j' \in [1, J] \\ \arg\max_{j \in [1,J]} g_{\perp,j}(\boldsymbol{x}) \quad \text{otherwise.} \end{cases}$$

Our proof of the soundness of Equation 8 follows the same approach that Mozannar and Sontag (2020) employed—specifically, a reduction to cost-sensitive learning.

Theorem 3.1. *Ψ_{SM}^J (Equation 8) is a convex (in g), calibrated surrogate loss for the $0-1$ multi-expert learning to defer loss (Equation 6).*

The complete proof is in Appendix A.2. The result guarantees that the minimizers $g_1^*, \ldots, g_K^*, g_{\perp,1}^*, \ldots, g_{\perp,J}^*$ correspond to the Bayes optimal classifier and rejector given in Equation 7.

3.2 One-vs-All Surrogate Loss

We next turn to the OvA surrogate loss. Let the label space $\mathcal{Y}^\perp$ and the functions $g_1, \ldots, g_K, g_{\perp,1}, \ldots, g_{\perp,J}$ be defined just as above for the softmax case. The OvA-based multi-expert L2D surrogate is then:

$$\Psi_{\mathrm{OvA}}^J(g_1, \ldots, g_K, g_{\perp,1}, \ldots, g_{\perp,J}; \boldsymbol{x}, y, m_1, \ldots, m_J) = \phi[g_y(\boldsymbol{x})] + \sum_{y' \in \mathcal{Y}, y' \neq y} \phi[-g_{y'}(\boldsymbol{x})] + \sum_{j=1}^{J} \phi[-g_{\perp,j}(\boldsymbol{x})] + \sum_{j=1}^{J} \mathbb{I}[m_j = y]\,(\phi[g_{\perp,j}(\boldsymbol{x})] - \phi[-g_{\perp,j}(\boldsymbol{x})]) \tag{9}$$

where $\phi : \{\pm 1\} \times \mathbb{R} \mapsto \mathbb{R}_+$ is again a binary surrogate loss. The classifier and rejector are computed exactly as in the softmax case.

We cannot construct our consistency proof in the same direct manner used in the softmax case. Like Verma and Nalisnick (2022), we proceed by the method of *error correcting output codes* (Dietterich and Bakiri, 1995; Langford et al., 2005; Allwein et al., 2001; Ramaswamy et al., 2014), a general technique for reducing multiclass problems to multiple binary problems. We prove the consistency of Ψ_{OvA}^J by way of the following two results.

Theorem 3.2. *For a strictly proper binary composite loss ϕ with a well-defined continuous inverse link function γ^{-1}, Ψ_{OvA}^J (Equation 9) is a calibrated surrogate for the $0-1$ multi-expert learning to defer loss (Equation 6).*

The complete proof is in Appendix A.3. Assuming *minimizability* (Steinwart, 2007)—i.e. that our hypothesis class is sufficiently large (all measurable functions)—the calibration result from Theorem 3.2 implies consistency.

Corollary 3.3. *Assume that $g \in \mathcal{F}$, where $\mathcal{F}$ is the hypothesis class of all measurable functions. Minimizability (Steinwart, 2007) is then satisfied for ψ^J_{OvA}, and it follows that ψ^J_{OvA} is a consistent surrogate for the $0-1$ multi-expert learning to defer loss (Equation 6).*

Thus, the minimizers of ψ^J_{OvA} (over all measurable functions) agree with the Bayes optimal classifier and rejector (Equation 7).

3.3 Inconsistency of Mixture of Experts

While we are the first to propose a consistent surrogate loss, previous work has proposed a *mixture of experts* (MoE) approach to multi-expert L2D. Hemmer et al. (2022) formulated the following model of the probability of the label under the whole team (of J experts and one classifier):

$$p\left(\mathrm{y}|\mathbf{x}, \mathrm{m}_1, \ldots, \mathrm{m}_J; \boldsymbol{\theta}_w, \boldsymbol{\theta}_h\right) = \\ w_0(\mathbf{x}; \boldsymbol{\theta}_w) \cdot p(\mathrm{y}|\mathbf{x}; \boldsymbol{\theta}_h) + \sum_{j=1}^{J} \mathbb{I}[\mathrm{m}_j = \mathrm{y}] \cdot w_j(\mathbf{x}; \boldsymbol{\theta}_w)$$

where $p(\mathrm{y}|\mathbf{x}; \boldsymbol{\theta}_h)$ denotes the classifier's probability. The function $\mathbf{w}(\mathbf{x}; \boldsymbol{\theta}_w) \in \triangle^{J+1}$—where $\triangle^{J+1}$ is the $(J+1)$-dimensional simplex—defines the mixture weights. w_0 assigns weight to the classifier, and w_j for $j \in [1, J]$ denotes the weight given to the jth expert. At test time, the index of the maximum weight determines to which downstream decision maker to assign responsibility. Hemmer et al. (2022) fit this MoE model using the negative log-likelihood of $p(\mathrm{y}|\mathbf{x}, \mathrm{m}_1, \ldots, \mathrm{m}_J; \boldsymbol{\theta}_w, \boldsymbol{\theta}_h)$; denote their loss $L_{\text{MoE}}(\boldsymbol{\theta}_w, \boldsymbol{\theta}_h)$. In Appendix A.4, we show that $L_{\text{MoE}}(\boldsymbol{\theta}_w, \boldsymbol{\theta}_h)$ is inconsistent. We provide a full discussion of related work in Section 6.

4 CONFIDENCE CALIBRATION

We next turn to the *calibration* (Dawid, 1982) properties of multi-expert L2D. While training with a consistent loss should produce models that are well-calibrated, previous work on the single-expert setting found that the underlying parameterizations can strongly influence calibration in practice. Specifically, Verma and Nalisnick (2022) show that the softmax formulation's estimators can be unbounded, resulting in 'probability' estimates above one. As for the calibration of the classifier, Verma and Nalisnick (2022) found that there is to systemic issue and can be improved with standard post-hoc techniques like temperature scaling (Kull et al., 2019), if necessary. Their findings also apply to the multi-expert scenario, and thus we consider only the rejector going forward.

We are particularly interested in the rejector's ability to estimate $\mathbb{P}(\mathrm{m}_j = \mathrm{y}|\boldsymbol{x})$, the conditional probability that the jth expert is correct. If the L2D system says that $\mathbb{P}(\mathrm{m}_j = \mathrm{y}|\boldsymbol{x}_0) = 0.7$, then the jth expert should be correct 70% of the time for inputs very similar to $\boldsymbol{x}_0$. This quantity is crucial not only for the system's ability to correctly defer but is also useful for interpretability and safety—to quantify what the model thinks that the human knows.

We next define the relevant notion of calibration. For an estimator of expert correctness $t(\boldsymbol{x}) : \mathcal{X} \mapsto (0, 1)$, we call t *calibrated* if, for any confidence level $c \in (0, 1)$, the actual proportion of times the expert is correct is equal to c:

$$\mathbb{P}(\mathrm{m} = \mathrm{y} \mid t(\boldsymbol{x}) = c) = c. \tag{10}$$

This statement should hold for all possible instances $\boldsymbol{x}$ with confidence c. Since expert correctness is a binary classification problem, distribution calibration, confidence calibration, and classwise calibration all coincide (Vaicenavicius et al., 2019). We can measure the degree of calibration using *expected calibration error* (ECE). In this case, the relevant ECE is defined as

$$\text{ECE}(t) = \mathbb{E}_{\mathbf{x}}|\mathbb{P}\left(\mathrm{m} = \mathrm{y} \mid t(\mathbf{x}) = c\right) - c|,$$

where $\mathbb{E}_{\mathbf{x}}$ is usually approximated with samples.

4.1 Softmax Parameterization

For the softmax formulation, the estimator of the probability that the jth expert is correct can be derived as follows; see Appendix A.2 (Equation 26). The Bayes optimal functions $g^*_1, \ldots, g^*_{\perp,J}$ have the following relationship with the underlying probability of expert correctness:

$$\frac{\mathbb{P}(\mathrm{m}_j = \mathrm{y}|\boldsymbol{x})}{1 + \sum_{j'=1}^{J} \mathbb{P}(\mathrm{m}_{j'} = \mathrm{y}|\boldsymbol{x})} = \underbrace{\frac{\exp\{g^*_{\perp,j}\}}{\sum_{y' \in \mathcal{Y}^\perp} \exp\{g^*_{y'}(\boldsymbol{x})\}}}_{t^*_{\perp,j}(\boldsymbol{x})}. \tag{11}$$

Denote the RHS of Equation 11 as $t^*_{\perp,j}(\boldsymbol{x})$. Since we have J equations, one for each expert, we can uniquely solve for $\mathbb{P}(\mathrm{m}_j = \mathrm{y}|\boldsymbol{x})$ as:

$$\mathbb{P}(\mathrm{m}_j = \mathrm{y}|\boldsymbol{x}) = \frac{t^*_{\perp,j}(\boldsymbol{x})}{1 - \sum_{j'=1}^{J} t^*_{\perp,j'}(\boldsymbol{x})}. \tag{12}$$

Equation 12 exhibits the same pathology as the single expert setting: it is unbounded from above. For $t_{\perp,j}(\boldsymbol{x}) > 0$, as $\sum_{j'=1}^{J} t_{\perp,j'}(\boldsymbol{x})$ approaches one, the estimate of $\mathbb{P}(\mathrm{m}_j = \mathrm{y}|\boldsymbol{x})$ will go to infinity. Moreover, the estimator for the jth expert depends on the estimators for the other experts. Thus, if one $t_{\perp,j}(\boldsymbol{x})$ is mis-calibrated, this error will likely propagate to the other estimators.

4.2 One-vs-All Parameterization

For the OvA formulation, the probability that the expert is correct is directly modeled by the jth deferral function. For the logistic binary loss ϕ, we have:

$$\begin{aligned}\mathbb{P}(\mathrm{m}_j = \mathrm{y}|\boldsymbol{x}) &= \phi\left(g^*_{\perp,j}(\boldsymbol{x})\right) \\ &= \frac{1}{1+\exp\{-g^*_{\perp,j}(\boldsymbol{x})\}}.\end{aligned} \tag{13}$$

This estimator has the correct range of $(0, 1)$ for any setting of $g_{\perp,j} \in \mathbb{R}$. Moreover, there is no dependence across expert deferral functions $g_{\perp,1}, \ldots, g_{\perp,J}$. We expect these properties to result in better calibration in practice.

5 ENSEMBLING EXPERTS WITH CONFORMAL INFERENCE

Multi-expert L2D, as defined above, operates by selecting just one expert upon deferral. This approach is sensible if querying each expert results in an independent expense (such as a consulting fee). However, in other settings, the cost incurred by deferring may just be that of time and efficiency (i.e. a lack of automation). In this case, the cost of querying additional experts would be negligible; for example, we could send multiple experts simultaneous messages asking for their decisions. Given the estimators of $\mathbb{P}(\mathrm{m}_j = \mathrm{y}|\boldsymbol{x})$ presented in the previous section, it is then natural to ask how we might ensemble experts according to these estimates of correctness. Below we present a methodology based on *conformal inference* for obtaining dynamic, minimal ensembles of experts.

Conformal Inference *Conformal inference* (CI) (Shafer and Vovk, 2008) constructs a confidence interval (or set) for predictive inference. In the traditional multiclass classification setting, given a new observation $\boldsymbol{x}_{n+1}$, we wish to determine the correct associated label $\mathrm{y}_{n+1} = y^*_{n+1}$, where y^*_{n+1} denotes the true class label. CI allows us to construct a distribution-free confidence set $\mathsf{C}(\boldsymbol{x}_{n+1})$ that will cover the true label with *marginal* probability $1-\alpha$:

$$\mathbb{P}\left(y^*_{n+1} \notin \mathsf{C}(\boldsymbol{x}_{n+1})\right) \;\leq\; \alpha \quad \forall\, \mathbb{P} \in \mathfrak{P}$$

where $\mathfrak{P}$ represents the space of all distributions—hence the 'distribution-free' quality. Denote the test statistic as $S(\boldsymbol{x}, y; \mathcal{D})$. It is known as a *non-conformity* function: a higher value of S means that $(\boldsymbol{x}, y)$ is less conforming to the distribution represented (empirically) by $\mathcal{D}$. Despite this guarantee, CI is only as good as its test statistic in practice. For instance, the marginal coverage is naively satisfied if we construct the set randomly by setting $\mathsf{C}(\boldsymbol{x}) = \mathcal{Y}$ with probability $1-\alpha$ and returning the empty set otherwise. CI is implemented by calculating the non-conformity function on a validation set and computing the empirical $1-\alpha$ quantile (with a finite sample correction). At test time, elements are added to the set until the non-conformity function passes the previously-computed quantile.

Conformal Sets of Experts We propose applying CI to perform uncertainty quantification for the experts. Thus, here, $\mathsf{C}(\boldsymbol{x})$ represents a set of experts. Firstly, we assume there is a best expert: for a new observation $\boldsymbol{x}_{n+1}$, let j^*_{n+1} denote the best expert such that

$$\mathbb{P}(\mathrm{m}_{j^*_{n+1}} = \mathrm{y}|\boldsymbol{x}_{n+1}) > \mathbb{P}(\mathrm{m}_e = \mathrm{y}|\boldsymbol{x}_{n+1})\; \forall e \neq j^*_{n+1}.$$

We would then like to construct a set such that j^*_{n+1} is covered with marginal probability $1-\alpha$:

$$\mathbb{P}\left(j^*_{n+1} \notin \mathsf{C}(\boldsymbol{x}_{n+1})\right) \;\leq\; \alpha \quad \forall\, \mathbb{P} \in \mathfrak{P}$$

where $\mathsf{C}(\boldsymbol{x}_{n+1})$ again denotes the conformal set and $\mathfrak{P}$ is the same as above. The set will have a dynamic size that changes with $\boldsymbol{x}$, ensuring our ensemble makes efficient use of expert queries. Unlike in most applications of CI, we can use the procedure to form an ensemble by aggregating the predictions of all experts in the set.

Naive Statistic We start by adapting a score function from multiclass classification. Let $s_j(\boldsymbol{x})$ denote the estimator that the jth expert is correct. For the softmax case, $s_j(\boldsymbol{x}) = t_{\perp,j}(\boldsymbol{x})/(1 - \sum_{j'} t_{\perp,j'}(\boldsymbol{x}))$ (Equation 12), and for OvA, $s_j(\boldsymbol{x}) = \phi(g_{\perp,j}(\boldsymbol{x}))$ (Equation 13). Let $\pi_1, \ldots, \pi_J$ denote the indices for a descending ordering of the estimators $s_j(\boldsymbol{x})$, i.e. s_{π_1} is the expert who has the best chance of being correct (according to the rejector). The resulting non-conformity function and test statistic are:

$$S\left(\boldsymbol{x}, y, m_1, \ldots, m_J; \mathcal{D}\right) = \sum_{e=1}^{E} s_{\pi_e}(\boldsymbol{x}) \tag{14}$$

where π_E is the index of the expert who has the lowest score s_{π_E} of all *correct experts* ($m = y$). This expression means that we will keep adding the correctness scores s in descending order until we include all experts who correctly predict the given instance. Hence $E = J$ only when all experts are correct and $E < J$ otherwise.

Regularized Statistic A problem with the statistic above is that multiple experts can be correct, resulting in noise that obscures the identity of the best expert. In the experiments, we show that this 'naive' statistic is not robust to noise, resulting in inflated set sizes (which are sometimes vacuous). Similar problems are discussed by Angelopoulos et al. (2020). To address this issue, we employ *conformal risk control* (Angelopoulos et al., 2022) to directly control the false negative rate. We create regularized prediction sets as follows:

$$C_{\lambda_\alpha}(\boldsymbol{x}) = \{j : s_j(\boldsymbol{x}) + \beta\left(s_j(\boldsymbol{x}) - \kappa\right) > 1 - \lambda_\alpha\} \tag{15}$$

where β and κ are the parameters of the regularization and λ_α is chosen to have $1-\alpha$ coverage guarantees. In Appendix

B, we describe λ_α and how we choose the regularization parameters. The general idea is to choose κ so that confidences lower than this threshold can happen with probability at most α. We choose β to optimize the size of the sets.

6 RELATED WORK

Learning to Defer *Learning to defer* (L2D) was proposed in its modern form by Madras et al. (2018). Yet classifiers with an option to reject or abstain date back at least to Chow (1957). There have been two primary approaches to making the rejection decision: confidence-based (Bartlett and Wegkamp, 2008; Yuan and Wegkamp, 2010; Jiang et al., 2018; Grandvalet et al., 2009; Ramaswamy et al., 2018; Ni et al., 2019) and model-based (Cortes et al., 2016a,b). Our work, due to it using a parameterized rejector, falls into the latter. The theoretical properties of the classifier-rejector approach have been well-studied for binary classification (Cortes et al., 2016a,b). The theory for multiclass classification was first developed by Ni et al. (2019) and Charoenphakdee et al. (2021). Mozannar and Sontag (2020) then built upon this work to establish the first consistent surrogate loss for multiclass L2D. Previous L2D extensions did not come with consistency guarantees (Raghu et al., 2019; Wilder et al., 2020; Pradier et al., 2021; Okati et al., 2021; Liu et al., 2022). Verma and Nalisnick (2022) proposed the second provably consistent surrogate for multiclass L2D based on a one-vs-all formulation. Charusaie et al. (2022) further studied the L2D optimization problem, proving results for complementarity and active learning. Our work extends Mozannar and Sontag (2020)'s and Verma and Nalisnick (2022)'s results to the multi-expert setting—for which no one has yet to propose a consistent surrogate loss.

Calibration in L2D Verma and Nalisnick (2022) motivate their OvA surrogate from the standpoint of calibration and thus is the only other work that studies the confidence calibration of L2D systems. We extend their work to the multi-expert setting. Calibration has received much attention of late in the wider machine learning literature (Guo et al., 2017; Kull et al., 2019; Vaicenavicius et al., 2019; Gupta and Ramdas, 2022). The dominant methodology is to apply post-hoc calibration: fitting additional parameters on validation data to re-calibrate the formerly mis-calibrated model. These methods could potentially be applied here—such as, by adding a temperature parameter to the per-expert terms in the OvA loss—but we are primarily interested in the native, 'out-of-the-box' calibration properties of the losses.

Multi-Expert Models There have been several works that use models to improve the decision making of multiple experts (Benz and Rodriguez, 2022; Straitouri et al., 2022) and to fuse decisions from models and humans (Keswani et al., 2021; Kerrigan et al., 2021). As mentioned in Section 3.3, Hemmer et al. (2022) proposed the only existing model for multi-expert L2D. Yet their approach does not have any supporting theoretical guarantees, such as consistency (like ours). Keswani et al. (2021) also proposed an MoE-based model but not for the standard L2D setting that we consider. Rather they allow for responsibility to be passed to multiple downstream sources—specifically, to any of the 2^{J+1} possible sets involving the experts and/or model.

Conformal Inference for Human-AI Collaboration We know of two works that have used CI for some form of human-AI collaboration. Straitouri et al. (2022) apply CI to a classifier and then pass the prediction set to a human to make the final decision. Babbar et al. (2022) study a similar work flow (apply CI then pass to a human) and also propose applying CI only to non-deferred samples, which results in smaller set sizes. No previous work has applied CI to obtain sets of experts.

7 EXPERIMENTS

Our experimental setup closely follows that of Verma and Nalisnick (2022)—but extended to multiple experts. For all runs, we report the mean and standard error across 3 random seeds. We perform three types of experiments. In the first, we check the system accuracy of the derived consistent surrogate losses in three consequential tasks (Subsection 7.1): galaxy classification, skin lesion diagnosis, and hate speech detection. We find that the OvA-trained model often outperforms both the softmax variant and MoE baseline. Secondly, we investigate the confidence calibration properties of the surrogates (Subsection 7.2). As hypothesized, the OvA loss results in less calibration error on both simulated and real data (possibly explaining its superior accuracy). Lastly, we investigate the efficacy of our conformal ensembling procedure (Subsection 7.3). For the naive statistic, the OvA loss' superior calibration results in appropriately smaller sets. For the regularized statistic, both losses perform equally well. Our implementations are publicly available at https://github.com/rajevv/Multi_L2D.

7.1 Overall System Accuracy

Data Sets We report the overall system accuracy for three real-world data sets: `HAM10000` (Tschandl et al., 2018) for skin lesions diagnosis, `Galaxy-Zoo` (Bamford et al., 2009) for galaxy classification, and `HateSpeech` (Davidson et al., 2017) for hate speech detection. The train-validation-test split is $60\% - 20\% - 20\%$. Following Verma and Nalisnick (2022), we down-sample `Galaxy-Zoo` to $10,000$ instances.

Models We use a 34-layer residual network (ResNet34) and a 50-layer residual network (ResNet50) as base models for `HAM10000` and `Galaxy-Zoo` respectively. For `HateSpeech`, we use a 100-dimensional *fasttext* (Joulin

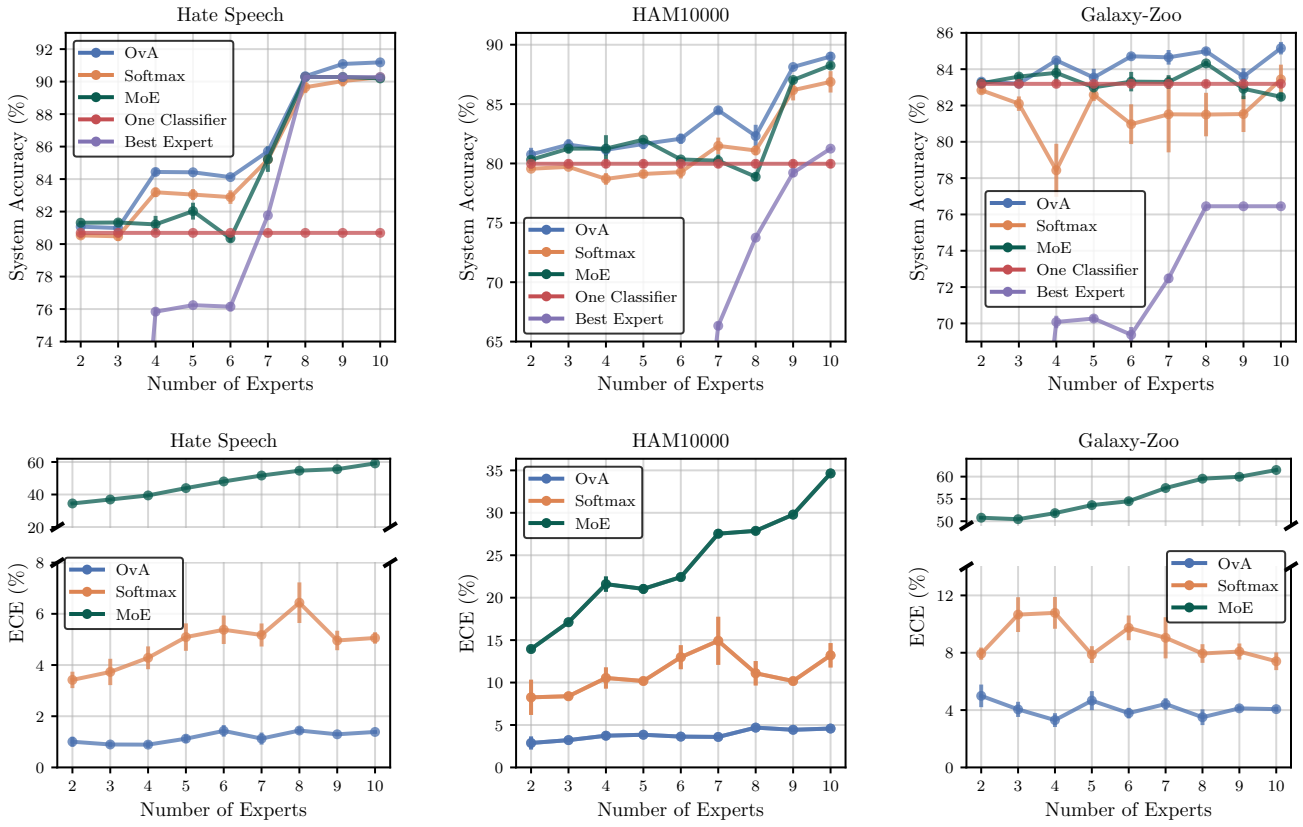


Figure 1: *System Accuracy and Calibration.* The figures above report the system accuracy (top row) and calibration error (bottom row) as experts of increasing ability are added (from 2 to 10). One classifier (red), best expert (purple), and a mixture of experts (green) (Hemmer et al., 2022) serve as baselines. We see that the OvA-trained model (blue) performs well in every case. On the other hand, the softmax-trained model (orange) falls below the one classifier baseline for both `HAM10000` and `GalaxyZoo`.

et al., 2016) representation of each input and a ConvNet (Kim, 2014) as the base model. We refer the reader to Verma and Nalisnick (2022) for more details on the training and hyperparameter selection, as we follow their setup.

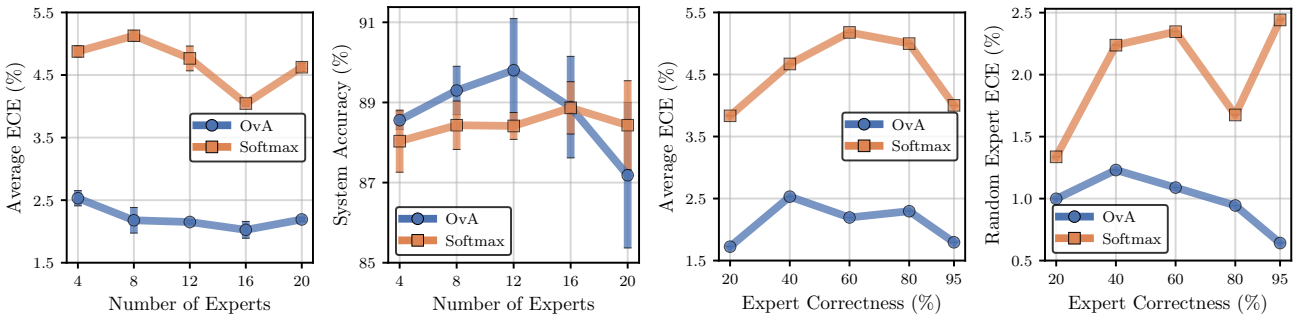
Experimental Setup We train the systems with an increasing number of experts, ranging from 2 to 10. See Appendix C for the details of how we simulate the experts. For each run, we enlarge the pool by adding increasingly accurate experts, and this process is repeated 3 times with different random seeds. We keep the base model fixed across these runs except for the additional output dimensions required by the expanded expert pool. Ideally, the L2D systems should exhibit strictly increasing accuracy due to adding experts of increasing quality. We compare our models against three baselines: one classifier, the best expert, and Hemmer et al. (2022)'s MoE.

Results The top row of Figure 1 reports the mean and standard error of the system accuracy as the number of experts increases. While the OvA, softmax, and MoE models perform comparably on `HateSpeech` (left), OvA's performance (blue) is notably better on `HAM10000` (center) and `Galaxy-Zoo` (right) as its accuracy never falls below the one classifier baseline (red), while the others' accuracies do.

7.2 Confidence Calibration

In Section 4, we found that the two surrogates have very different estimators of $\mathbb{P}\left(\mathrm{m}_j = \mathrm{y}_i | \mathbf{x}_i\right)$, the probability that the jth expert is correct. We now test if these theoretical differences have consequences for practice. To ensure ECE is well-defined for the softmax loss, we cap any confidences greater than 1 at 1. In addition to reporting calibration for the preceding experiment (system accuracy), we also perform simulations using the standard splits of `CIFAR-10` (Krizhevsky, 2009). We use a 28-layer wide residual network (Zagoruyko and Komodakis, 2016), following Verma and Nalisnick (2022).

Simulation #1: Increasing Experts We perform a simulation to see how the methods perform under an increasing



(a) ECE as Experts Increase (b) Acc. as Experts Increase (c) ECE for Improving Experts (d) ECE of Random Expert

Figure 2: *Confidence Calibration Simulations.* The figures above report the results for confidence calibration simulations performed on `CIFAR-10`. The first and second subfigures show calibration error (a) and system accuracy (b) as the number of experts increases from 4 to 20. We see that the OvA formulation (blue) has better calibration across all runs, but this translate to better accuracy only for 16 or fewer experts. The third and forth subfigures show calibration error as experts' abilities increase (a) and when one expert is kept at random chance (b). OvA (blue) shows better calibration in both metrics.

number of experts. We generate a synthetic expert with a correctness probability of 70% over the first five classes and random across all other classes. We then replicate that expert and add it to the expert pool, ranging from 4 to 20 total experts. Figure 2a reports the average ECE across experts as the pool increases. The OvA method (blue) is roughly stable at about 2% ECE as experts are added. The softmax method (orange) has roughly double the ECE ($\sim 4.5\%$). In Figure 2b, we report the overall system accuracy to see if these calibration differences have an effect. We see some positive effects, with OvA (blue) having a better accuracy for 12 and fewer experts. However, the softmax (orange) has the best accuracy at 20 experts, despite its calibration still being worse.

Simulation #2: Expert Dependence We next perform a simulation to see how calibration error can propagate across the estimators. We simulate four experts with one always being random and the other three having a probability of correctness that increases from 20% to 95% on the first five classes (random for others). We hypothesize that the softmax's ECE for the random expert will *increase* when the probability of correctness for the other three experts increases due to the tied parameterization. Figures 2c and 2d report the results, with the former reporting average ECE and the latter the ECE of just the random expert. Firstly, from Figure 2c, we see that again the OvA method is better calibrated across all experimental settings. Then from Figure 2d, we see that our hypothesis is confirmed: OvA (blue) is able to model the random expert well no matter the other experts' abilities, but the softmax (orange) is not. The softmax's ECE increases almost in-step with the expert correctness, except for some cancellation effect happening at 80%. This is clearly an undesirable behavior from the standpoint of safety since any ECE above zero means that the system is reporting that the expert is better than random and thus misleading the user.

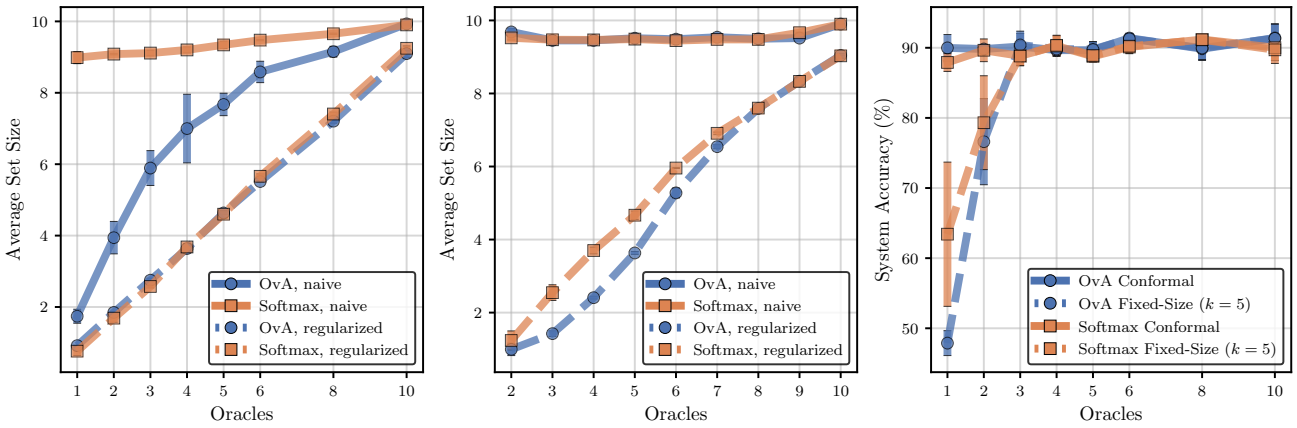
Hate Speech, HAM10000, and Galaxy-Zoo Lastly we report the calibration error of the models trained for the experiments reported in Section 7.1. The results are in the bottom row of Figure 1. The trend we observe in the `CIFAR-10` simulations is also observed here, with OvA (blue) having the best calibration. This may explain why OvA has the best system accuracy. Unsurprisingly, the MoE has extremely poor calibration, which is likely due to its inconsistent optimization objective which allows for sub-optimal models (as we prove in Proposition A.4).

7.3 Conformal Ensembles

Lastly, we study our proposal of using CI to ensemble multiple experts. We first analyze the two proposed statistics, demonstrating the regularized version's superior ability to recover the experts who are oracles. We then report the downstream effect on the overall system accuracy, comparing performance to that of a fixed-size ensemble of experts.

Experts and Setup We experiment with two settings on `CIFAR-10`, each with 10 total experts. In the first (*no noise*), we synthesize experts such that they are an oracle on an increasing subset of the classes and guaranteed to be wrong on the classes not in that set. In the second (*with noise*), the experts are oracles in the same way but now have a (uniformly) random chance of being correct for the non-oracle classes. The theory of CI guarantees that the sets *marginally* cover all oracle experts. Yet, ideally, we wish the sets to contain *only* the experts who are oracles. We use $\alpha = 0.1$ in all experiments.

Expert Identification The CI results are reported in Figures 3a and 3b. The number of oracles is on the y-axis, and the average set size is on the x-axis. Optimal performance would be the $y = x$ line. The results for the no-noise setting are reported in Figure 3a. The naive statistic (solid lines)



(a) Set Size, No Noise (b) Set Size, with Noise (c) CIFAR-10: Sys. Accuracy

Figure 3: *Conformal Sets of Experts.* The figures above report our analysis of the two statistics proposed in Section 5. Subfigures (a) and (b) show the ability of the two statistics to select the correct number of experts as the number of oracle experts increases—so optimal performance is the $y = x$ line. Subfigure (a) reports the no-noise setting (so experts are either perfectly incorrect or correct), and we see that the naive statistic (solid lines) overestimates the set size. The problem is even worse in Subfigure (b). However, the regularized statistic (dashed lines) is able to do well in both cases. Subfigure (c) shows how ensembling the set affects system accuracy. The conformal approach is able to out-perform a fixed size of 5 experts for a small number of oracles and is equivalent at higher numbers.

considerably inflates the set size for both softmax and OvA. Yet OvA is much closer to $y = x$, which suggests superior calibration leads to better CI. The performance of the regularized statistic is shown by the dashed lines. Both softmax and OvA perform nearly perfectly. Figure 3b reports the with-noise setting, and we find that the naive statistic performs terribly for both losses. The regularized statistic, on the other hand, performs well for both softmax and OvA. Softmax demonstrates slight superiority for $2 - 5$ oracles.

Overall System Accuracy We next investigate using the conformal set as an ensembling strategy. Upon deferral, we use majority voting across the set to generate the final prediction. We compare this with the baseline of using a fixed ensemble size—specifically, the top five ranked experts. We show the results for `CIFAR-10` in Figure 3c. The crucial settings are for one and two oracles since using a fixed ensemble size is guaranteed to fail—as is confirmed by the plot ($< 80\%$ accuracy). We see that the conformal ensembles are clearly superior here, achieving around 90% accuracy. For three or more oracles, both methods have equal performance. This is expected since only three oracles are needed to form a correct majority. We emphasize that CI's *adaptivity* is highly desirable so that the best experts can be identified with transparency and queried efficiently.

8 CONCLUSIONS

We have extended the L2D framework to support multiple experts. We proposed two optimization objectives and proved that they are both consistent. Our proposed optimization objectives are simple to use in practice and could be embedded into any empirical risk minimization framework. Additionally, we also studied their potential to be confidence calibrated, showing that the softmax-based objective can result in mis-calibrated models in practice. Lastly, we considered a principled procedure for selecting *minimal* sets of experts to ensemble. For future work, we aim to improve the data efficiency of our method by extending the active learning results of Charusaie et al. (2022) to the multi-expert setting. Additionally, recent work by Narasimhan et al. (2022) has shown that L2D models can be prone to under-fitting when querying experts incurs additional cost and propose a post-hoc correction for surrogate losses. Extending this methodology to the multi-expert setting may generally improve our results.

Acknowledgements

This publication is part of the project *Continual Learning under Human Guidance* (VI.Veni.212.203), which is financed by the Dutch Research Council (NWO). This work has also been partly supported by the Spanish government (AEI/MCI) under grant PID2021-123182OB-I00, by Comunidad de Madrid under grants IND2022/TIC-23550 and IntCARE-CM, and by the European Union (FEDER) and the European Research Council (ERC) through the European Union's Horizon 2020 research and innovation program under grant 714161. The work by Daniel Barrejon has been additionally funded by the Spanish Ministerio de Educaciòn, Cultura Deporte, grant FPU19/02681. This work was carried out on the Dutch national e-infrastructure with the support of the SURF Cooperative.

References

Zeynep Akata, Dan Balliet, Maarten De Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, et al. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(08):18–28, 2020.

Erin L. Allwein, Robert E. Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1: 113–141, 2001.

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2020.

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

Varun Babbar, Umang Bhatt, and Adrian Weller. On the utility of prediction sets in human-ai teams. In *International Joint Conference on Artificial Intelligence*, 2022.

Steven P. Bamford, Robert C. Nichol, Ivan K. Baldry, Kate Land, Chris J. Lintott, Kevin Schawinski, Anže Slosar, Alexander S. Szalay, Daniel Thomas, Mehri Torki, Dan Andreescu, Edward M. Edmondson, Christopher J. Miller, Phil Murray, M. Jordan Raddick, and Jan Vandenberg. Galaxy Zoo: the dependence of morphology and colour on environment*. *Monthly Notices of the Royal Astronomical Society*, 393(4):1324–1352, 2009.

Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(59):1823–1840, 2008.

Nina Corvelo Benz and Manuel Gomez Rodriguez. Counterfactual inference of second opinions. In *Conference on Uncertainty in Artificial Intelligence*, 2022.

Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. Classification with rejection based on cost-sensitive classification. In *International Conference on Machine Learning*, 2021.

Mohammad-Amin Charusaie, Hussein Mozannar, David Sontag, and Samira Samadi. Sample efficient learning of predictors that complement humans. In *International Conference on Machine Learning*, 2022.

C. K. Chow. An optimum character recognition system using decision functions. *IRE Trans. Electron. Comput.*, 6:247–254, 1957.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, 2016a.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Boosting with abstention. In *Advances in Neural Information Processing Systems*, 2016b.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *International AAAI Conference on Web and Social Media*, 2017.

A Philip Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. Hybrid intelligence. *Business & Information Systems Engineering*, 61(5):637–643, 2019.

Thomas G. Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2(1):263–286, 1995.

Doris Fay, Carol Borrill, Ziv Amir, Robert Haward, and Michael A West. Getting the most out of multidisciplinary teams: A multi-sample study of team innovation in health care. *Journal of Occupational and Organizational Psychology*, 79(4):553–567, 2006.

Yves Grandvalet, Alain Rakotomamonjy, Joseph Keshet, and Stéphane Canu. Support vector machines with a reject option. In *Advances in Neural Information Processing Systems*, 2009.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Chirag Gupta and Aaditya Ramdas. Top-label calibration and multiclass-to-binary reductions. In *International Conference on Learning Representations*, 2022.

Patrick Hemmer, Sebastian Schellhammer, Michael Vössing, Johannes Jakubik, and Gerhard Satzger. Forming Effective Human-AI Teams: Building Machine Learning Models that Complement the Capabilities of Multiple Experts. In *International Joint Conference on Artificial Intelligence*, 2022.

Heinrich Jiang, Been Kim, Melody Y. Guan, and Maya Gupta. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems*, 2018.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomás Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

Ece Kamar. Directions in hybrid intelligence: Complementing ai systems with human intelligence. In *International Joint Conference on Artificial Intelligence*, 2016.

Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. Combining human predictions with model probabilities via confusion matrices and calibration. *Advances in Neural Information Processing Systems*, 2021.

Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Towards unbiased and accurate deferral to multiple experts. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2021.

Yoon Kim. Convolutional neural networks for sentence classification. In *Conference on Empirical Methods in Natural Language Processing*, 2014.

Harold Kittler, H Pehamberger, K Wolff, and MJTIO Binder. Diagnostic accuracy of dermoscopy. *The lancet oncology*, 3(3):159–165, 2002.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *Technical Report*, 2009.

Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems*, 2019.

John Langford, Tti-Chicago, JI@hunch Net, and Alina Beygelzimer. Sensitive error correcting output codes. In *Conference on Learning Theory*, 2005.

Jessie Liu, Blanca Gallego, and Sebastiano Barbieri. Incorporating Uncertainty in Learning to Defer Algorithms for Safe Computer-Aided Diagnosis. *Scientific reports*, 12 (1):1–9, 2022.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, 2018.

Hussein Mozannar and David A. Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, 2020.

Harikrishna Narasimhan, Wittawat Jitkrittum, Aditya Krishna Menon, Ankit Singh Rawat, and Sanjiv Kumar. Post-hoc estimators for learning to defer to an expert. In *Advances in Neural Information Processing Systems*, 2022.

Chenri Ni, Nontawat Charoenphakdee, Junya Honda, and Masashi Sugiyama. On the calibration of multiclass classification with rejection. In *Advances in Neural Information Processing Systems*, 2019.

Nastaran Okati, Abir De, and Manuel Gomez-Rodriguez. Differentiable learning under triage. In *Advances in Neural Information Processing Systems*, 2021.

Melanie F Pradier, Javier Zazo, Sonali Parbhoo, Roy H Perlis, Maurizio Zazzi, and Finale Doshi-Velez. Preferential mixture-of-experts: Interpretable models that rely on human expertise as much as possible. *AMIA Summits on Translational Science Proceedings*, 2021:525, 2021.

Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*, 2019.

H. G. Ramaswamy, Ambuj Tewari, and Shivani Agarwal. Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12: 530–554, 2018.

Harish G. Ramaswamy, Balaji Srinivasan Babu, Shivani Agarwal, and Robert C. Williamson. On the consistency of output code based learning algorithms for multiclass learning problems. In *Conference on Learning Theory*, 2014.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12): 371–421, 2008.

Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26:225–287, 2007.

Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez Rodriguez. Provably improving expert predictions with conformal prediction. *arXiv preprint arXiv:2201.12006*, 2022.

Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. In *Advances in Neural Information Processing Systems*, 2021.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *Conference on Artificial Intelligence and Statistics*, 2019.

Rajeev Verma and Eric Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *International Conference on Machine Learning*, 2022.

Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. In *International Joint Conference on Artificial Intelligence*, 2020.

Ming Yuan and Marten Wegkamp. Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, 11(5):111–130, 2010.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, 2016.

A Proofs and Derivations

In this section, we provide proofs for the main results in the paper. We derive the Bayes optimal rule for L2D to multiple experts, and show that the surrogate losses proposed in the paper are consistent. Next, we show that the mixture of experts formulation (Hemmer et al., 2022) is not consistent. We continue the notation from the main paper. For simplicity, we do not worry about measure-theoretic considerations and assume that appropriate conditions hold that allow us to interchange summations and integrals, for example. We begin by giving a formal definition of what it means for a surrogate loss to be consistent.

Definition A.1. (Consistent Loss Function). A surrogate loss function $\psi : \mathcal{C} \times \mathcal{Y} \to \mathbb{R}_+$ operating in the surrogate space $\mathcal{C} \subseteq \mathbb{R}^K$ along with some suitable decoding function $g : \mathcal{C} \to \mathcal{Y}$ is said to be consistent if for all distributions $\mathcal{D}$, $\forall \epsilon > 0$, $\exists\, \delta > 0$ such that if

$$\left|\mathbb{E}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}}\left[\psi\left(\mathrm{y}, \mathbf{c}(\mathbf{x})\right)\right] - \inf_{\mathbf{u}\in\mathcal{C}} \mathbb{E}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}}\left[\psi\left(\mathrm{y}, \mathbf{u}\right)\right]\right| < \delta, \tag{16}$$

holds for a prediction function $h : \mathcal{X} \to \mathcal{C}$, $h(\mathbf{x}) := \mathbf{c}(\mathbf{x}) \in \mathbb{R}^K$, then it must hold that

$$\mathbb{P}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}}\left[g \circ h(\mathbf{x}) \neq \mathrm{y}\right] \leq \mathbb{P}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}}\left[h^*(\mathbf{x}) \neq \mathrm{y}\right] + \epsilon, \tag{17}$$

where $h^* : \mathcal{X} \to \mathcal{Y}$ is the Bayes optimal predictor.

A common example of one decoding function in machine learning is the $\arg\max$ function. Intuitively, consistency implies that the minimization of a surrogate loss $\psi(\cdot)$ results in a prediction function $h(\cdot)$ whose expected error converges to the Bayes risk.

A.1 Bayes Rule for Learning to Defer with Multiple Experts

We have J experts and a classifier, where the system either allows the classifier to make the final prediction or defers to one of the J experts. When the classifier makes the prediction, the system incurs loss $\ell_{\text{clf}}(\hat{\mathrm{y}}, \mathrm{y})$ where $\hat{\mathrm{y}} = h(\boldsymbol{x})$. When the system defers to the j^{th} expert, it incurs a loss $\ell_{\text{exp}}(\mathrm{m}_j, \mathrm{y})$. In what follows, we frame the learning to defer problem as a general classification problem, and aim to find a function $g : \mathcal{X} \to \hat{\mathcal{Y}} := \mathcal{Y} \cup \{\perp_1, \perp_2, \ldots, \perp_J\}$ and $|\hat{\mathcal{Y}}| = K + J$ with the minimum expected loss (also known as *risk*). We consider g as modeling a probabilistic decision rule $\delta(\hat{y}|\boldsymbol{x}) := [\delta(\hat{y} = 1|\boldsymbol{x}), \delta(\hat{y} = 2|\boldsymbol{x}), \ldots, \delta(\hat{y} = K + J|\boldsymbol{x})]$ where $\delta(\hat{y} = i|\boldsymbol{x})$ denotes the confidence in making the i^{th} decision for $\boldsymbol{x} \sim \mathbf{x}$. We write *risk* as follows:

$$\mathcal{R}_{\mathcal{D}}[\delta\left(\hat{y}|\mathbf{x}\right)] = \sum_{i=1}^{\mathrm{K}} \sum_{j=1}^{\mathrm{K+J}} \int_{\mathbf{x}} \delta(\hat{y} = j|\mathbf{x}) \ell\left(\hat{\mathrm{y}} = j, \mathrm{y} = i\right) \mathbb{P}\left(\mathrm{y} = i\right) \mathbb{P}\left(\mathbf{x}|\mathrm{y} = i\right) d\mathbf{x}, \tag{18}$$

where $\ell : (\hat{\mathrm{y}}, \mathrm{y}) \mapsto \mathbb{R}_+$ is a general loss function, i runs over the input label space, and j runs over the output prediction space (classifier and all the experts). We further expand the risk in Equation 18 based on the definition of the loss function in learning to defer

$$\begin{aligned}\mathcal{R}_{\mathcal{D}}[\delta\left(\hat{y}|\mathbf{x}\right)] = \sum_{i=1}^{\mathrm{K}} \int_{\mathbf{x}} \Bigg(& \sum_{j=1}^{\mathrm{K}} \delta(\hat{y}_j|\mathbf{x}) \ell_{\text{clf}}\left(j, i\right) \\ & + \sum_{j=\mathrm{K+1}}^{\mathrm{K+J}} \left(\sum_{m=1}^{\mathrm{K}} \delta(\hat{y}_j|\mathbf{x}) \ell_{exp}\left(m_j, \mathrm{y}\right) \mathbb{P}(m_j|\mathbf{x}, \mathrm{y} = i) \right) \Bigg) \mathbb{P}\left(\mathrm{y} = i\right) \mathbb{P}\left(\mathbf{x}|\mathrm{y} = i\right) d\mathbf{x},\end{aligned}$$

where we have used shorthand $\delta\left(\hat{y}_j|\mathbf{x}\right)$ to denote $\delta\left(\hat{y} = j|\mathbf{x}\right)$, and $\mathbb{P}\left(m_j|\mathbf{x}, \mathrm{y} = i\right) = \mathbb{P}\left(\mathrm{m}_j = m|\mathbf{x}, \mathrm{y} = i\right)$. Next, we denote:

$$\begin{aligned} w_{i,j} &= \ell_{\text{clf}}(j, i) \\ w_{i,\perp_j} &= \sum_{m=1}^{\mathrm{K}} \delta(\hat{y}_j|\mathbf{x}) \ell_{exp}\left(m_j, \mathrm{y}\right) \mathbb{P}(m_j|\mathbf{x}, \mathrm{y} = i)) \end{aligned}$$

Thus, $\mathcal{R}_{\mathcal{D}}[\delta\left(\hat{y}|\mathbf{x}\right)]$ can be written as:

$$\mathcal{R}_{\mathcal{D}}[\delta\left(\hat{y}|\mathbf{x}\right)] = \sum_{i=1}^{\mathrm{K}} \int_{\mathbf{x}} \left(\sum_{j=1}^{\mathrm{K}} \delta(\hat{y}_j|\mathbf{x}) w_{i,j} + \sum_{j=\mathrm{K}+1}^{\mathrm{K}+\mathrm{J}} \delta(\hat{y}_j|\mathbf{x}) w_{i,\perp_j} \right) \mathbb{P}\left(\mathrm{y}=i\right) \mathbb{P}\left(\mathbf{x}|\mathrm{y}=i\right) d\mathbf{x}.$$

Denote $w^*_{i,\perp} := \min_{j\in[J]}\{w_{i,\perp_j}\}$, we have

$$\mathcal{R}_{\mathcal{D}}[\delta\left(\hat{y}|\mathbf{x}\right)] \geq \sum_{i=1}^{\mathrm{K}} \int_{\mathbf{x}} \left(\sum_{j=1}^{\mathrm{K}} \delta(\hat{y}_j|\mathbf{x}) w_{i,j} + \sum_{j=\mathrm{K}+1}^{\mathrm{K}+\mathrm{J}} \delta(\hat{y}_j|\mathbf{x}) w^*_{i,\perp} \right) \mathbb{P}\left(\mathrm{y}=i\right) \mathbb{P}\left(\mathbf{x}|\mathrm{y}=i\right) d\mathbf{x}.$$

We also denote $\sum_{j=\mathrm{K}+1}^{\mathrm{K}+\mathrm{J}} \delta(\hat{y}_j|\mathbf{x})$ as $\delta(\hat{y}_\perp|\mathbf{x})$. Then the lower bound of $\mathcal{R}_{\mathcal{D}}[\delta\left(\hat{y}|\mathbf{x}\right)]$, denoted as $\bar{\mathcal{R}}_{\mathcal{D}}[\delta\left(\hat{y}|\mathbf{x}\right)]$, is

$$\bar{\mathcal{R}}_{\mathcal{D}}[\delta\left(\hat{y}|\mathbf{x}\right)] = \sum_{i=1}^{\mathrm{K}} \int_{\mathbf{x}} \left(\sum_{j=1}^{\mathrm{K}} \delta(\hat{y}_j|\mathbf{x}) w_{i,j} + \delta(\hat{y}_\perp|\mathbf{x}) w^*_{i,\perp} \right) \mathbb{P}\left(\mathrm{y}=i\right) \mathbb{P}\left(\mathbf{x}|\mathrm{y}=i\right) d\mathbf{x}.$$

Since $\sum_{j=1}^{\mathrm{K}} \delta(\hat{y}_j|\mathbf{x}) + \delta(\hat{y}_\perp|\mathbf{x}) = 1.0$ and $\int_{\mathbf{x}} \mathbb{P}(\mathbf{x}|\mathrm{y}=i)d\mathbf{x} = 1.0$, we follow Chow (1957) to decompose $\bar{\mathcal{R}}_{\mathcal{D}}[\delta\left(\hat{y}|\mathbf{x}\right)]$ in two terms:

$$\bar{\mathcal{R}}_{\mathcal{D}} = \bar{\mathcal{R}}^{\perp}_{\mathcal{D}} + \bar{\mathcal{R}}^{\delta}_{\mathcal{D}},$$

where

$$\begin{aligned}
\bar{\mathcal{R}}^{\perp}_{\mathcal{D}} &= \sum_{i=1}^{\mathrm{K}} \mathbb{P}(\mathrm{y}=i) \cdot w^*_{i,\perp}, \\
\bar{\mathcal{R}}^{\delta}_{\mathcal{D}} &= \int_{\mathbf{x}} \sum_{j=1}^{[\mathrm{K}]\cup\{\perp\}} \delta(\hat{y}_j|\mathbf{x}) Z_j(\mathbf{x}) d\mathbf{x}, \text{and} \\
Z_j(\mathbf{x}) &= \sum_{i=1}^{\mathrm{K}} \left(w_{i,j} - w^*_{i,\perp}\right) \mathbb{P}(\mathbf{x}) \mathbb{P}(\mathrm{y}=i|\mathbf{x}),\ j \in \{1,2,\ldots,\mathrm{K},\perp\}.
\end{aligned}$$

To elaborate, we simplify the problem from deferring to multiple experts to deferring to just the one expert with the minimum $w_{i,\perp_j}$ in obtaining the lower bound $\bar{\mathcal{R}}^{\perp}_{\mathcal{D}}$. We also observe that we have no control over $\bar{\mathcal{R}}^{\perp}_{\mathcal{D}}$. However, we can control $\bar{\mathcal{R}}^{\delta}_{\mathcal{D}}$ by controlling the decision rule δ. We have $Z_\perp(\mathbf{x}) = 0$, and also it holds that

$$\bar{\mathcal{R}}^{\delta}_{\mathcal{D}} \geq \int_{\mathbf{x}} \min_j \left[Z_j\left(\mathbf{x}\right)\right] d\mathbf{x},$$

where the equality holds iff $\delta\left(\hat{y}_k|\mathbf{x}\right) = 1.0$ for $k = \arg\min_j Z_j\left(\mathbf{x}\right)$. Thus, the optimal rule is to deterministically (i.e. with confidence 1.0) choose $k \in \{1,2,\ldots,\mathrm{K},\perp\}$ with the minimum $Z_j\left(\mathbf{x}\right)$. This means that choosing j for which the $Z_j(\mathbf{x})$ is the smallest minimizes the risk. Given that $Z_\perp = 0$, this means that the classifier predicts when the minimum $Z_j\left(\mathbf{x}\right)$ is negative. Thus, deferral happens when $Z_j(\mathbf{x})$ is positive for all j, i.e., the optimal rejection rule $r^*(\mathbf{x})$ is:

$$r^*(\mathbf{x}) = \mathbb{I}\left[Z_j(\mathbf{x}) \geq 0; \forall j \in \{1,\ldots,\mathrm{K}\}\right]. \tag{19}$$

This rejection rule is similar to the learning to defer to one expert. Given the definition of $Z_j(\mathbf{x})$, the optimal behavior to choose which expert to defer to is the one with minimum $w_{i,\perp j}$. We further simplify the optimal rejection rule in the following proposition.

Proposition A.2. *The Bayes optimal rejection rule for L2D with multiple experts is given as:*

$$r^*(\mathbf{x}) = \begin{cases} 1 & \textit{if } \mathbb{E}_{\mathrm{y}|\mathbf{x}}\left[\ell_{\textit{clf}}(\hat{\mathrm{y}},\mathrm{y})\right] \geq \min_{j\in J} \mathbb{E}_{\mathrm{y}|\mathbf{x}} \mathbb{E}_{\mathrm{m}|\mathbf{x},\mathrm{y}}\left[\ell_{\textit{exp}}\left(\mathrm{m}_j,\mathrm{y}\right)\right] \quad \forall \hat{\mathrm{y}} \in \mathcal{Y} \\ 0 & \textit{otherwise}. \end{cases} \tag{20}$$

Proof: The proof follows immediately from the definition of $r^*(\mathbf{x})$ in Equation 19.

In our work, we use the canonical 0-1 loss for both ℓ_{clf} and ℓ_{exp}. In this case, the rejection rule can trivially be written as in the following corollary.

Corollary A.3. *For a misclassification* 0-1 *loss, the optimal rejection rule is:*

$$r^*(\mathbf{x}) = \mathbb{I}\left[\max_{j\in J}\mathbb{P}(\mathrm{m}_j = \mathrm{y}|\mathbf{x}=\boldsymbol{x}) \geq \max_{y\in\mathcal{Y}}\mathbb{P}(\mathrm{y}=y|\mathbf{x}=\boldsymbol{x})\right], \tag{21}$$

where $\mathbb{P}(\mathrm{m}_j = \mathrm{y}|\mathbf{x}=\boldsymbol{x})$ is the expert's correctness probability for the j^{th} expert, and $\mathbb{P}(\mathrm{y}=y|\mathbf{x}=\boldsymbol{x})$ is the regular class probability.

To sum it up, the Bayes optimal rule is to compare the confidences of the experts and the classifier, and follow whosoever has the highest confidence. The rule is analogous to the single expert setting proved in Mozannar and Sontag (2020).

A.2 Proof of Theorem 3.1: Consistency of ψ^J_{SM}

Convexity of ψ^J_{SM} is immediately clear. We provide the proof for consistency below:

For simplicity, we denote

$$\begin{aligned}
-\log\left(\frac{\exp\{g_y(\boldsymbol{x})\}}{\sum_{y'\in\mathcal{Y}^\perp}\exp\{g_{y'}(\boldsymbol{x})\}}\right) &= \zeta_y(\boldsymbol{x}),\\
-\log\left(\frac{\exp\{g_{\perp,j}(\boldsymbol{x})\}}{\sum_{y'\in\mathcal{Y}^\perp}\exp\{g_{y'}(\boldsymbol{x})\}}\right) &= \zeta_{\perp,j}(\boldsymbol{x}).
\end{aligned} \tag{22}$$

Then, ψ^J_{SM} can be written as:

$$\phi^J_{\text{SM}}\left(g_1,\ldots,g_K,g_{\perp,1},\ldots,g_{\perp,J};\boldsymbol{x},y,m_1,\ldots,m_J\right) = \zeta_y(\boldsymbol{x}) + \sum_{j=1}^{J}\mathbb{I}[m_j=y]\cdot\zeta_{\perp,j}(\boldsymbol{x}). \tag{23}$$

We consider the pointwise risk $\mathcal{C}[\psi^J_{\text{SM}}]$ defined as:

$$\mathcal{C}[\psi^J_{\text{SM}}] = \mathbb{E}_{\mathrm{y}|\mathbf{x}=\boldsymbol{x}}\mathbb{E}_{\mathrm{m}|\mathbf{x}=\boldsymbol{x},\mathrm{y}=y}\left[\psi^J_{\text{SM}}(g_1,\ldots,g_K,g_{\perp,1},\ldots,g_{\perp,J};\boldsymbol{x},y,m_1,\ldots,m_J)\right], \tag{24}$$

where $\mathrm{m}|\mathbf{x}=\boldsymbol{x},\mathrm{y}=y$ is a compact representation for each $\mathrm{m}_j|\mathbf{x}=\boldsymbol{x},\mathrm{y}=y$. Our setup assumes that each m_j is independent. Denote $\eta_y(\boldsymbol{x}) = \mathbb{P}(\mathrm{y}=y|\mathbf{x}=\boldsymbol{x})$, we expand the expectations:

$$\begin{aligned}
\mathcal{C}[\psi^J_{\text{SM}}] &= \sum_{y\in\mathcal{Y}}\eta_y(\boldsymbol{x})\cdot\zeta_y(\boldsymbol{x}) + \sum_{j=1}^{J}\left(\sum_{y\in\mathcal{Y}}\eta_y(\boldsymbol{x})\sum_{m_j\in\mathcal{M}}\mathbb{P}(\mathrm{m}_j=m_j|\mathbf{x}=\boldsymbol{x},\mathrm{y}=y)\mathbb{I}[m_j=y]\cdot\zeta_{\perp,j}(\boldsymbol{x})\right).\\
\mathcal{C}[\psi^J_{\text{SM}}] &= \sum_{y\in\mathcal{Y}}\eta_y(\boldsymbol{x})\cdot\zeta_y(\boldsymbol{x}) + \sum_{j=1}^{J}\left(\sum_{y\in\mathcal{Y}}\eta_y(\boldsymbol{x})\sum_{m_j\in\mathcal{M}}\mathbb{P}(\mathrm{m}_j=y|\mathbf{x}=\boldsymbol{x},\mathrm{y}=y)\cdot\zeta_{\perp,j}(\boldsymbol{x})\right).\\
\mathcal{C}[\psi^J_{\text{SM}}] &= \sum_{y\in\mathcal{Y}}\eta_y(\boldsymbol{x})\cdot\zeta_y(\boldsymbol{x}) + \sum_{j=1}^{J}\mathbb{P}(\mathrm{m}_j=\mathrm{y}|\mathbf{x}=\boldsymbol{x})\cdot\zeta_{\perp,j}(\boldsymbol{x}).
\end{aligned}$$

Next, we consider the minimizer of $\mathcal{C}[\psi^J_{\text{SM}}]$. Since we have established convexity, we can analyze the minimizers of $\mathcal{C}[\psi^J_{\text{SM}}]$ by taking the partial derivatives w.r.t. $g_y\{\boldsymbol{x}\}$ and $g_{\perp,j}\{\boldsymbol{x}\}$ respectively and set them to 0.

Thus, w.r.t. $g_y\{\boldsymbol{x}\}$, we have

$$\frac{\partial\mathcal{C}[\psi^J_{\text{SM}}]}{\partial g_y\{\boldsymbol{x}\}} = 0 \implies \frac{\exp\{g_y(\boldsymbol{x})\}}{\sum_{y'\in\mathcal{Y}^\perp}\exp\{g_{y'}(\boldsymbol{x})\}} = \frac{\mathbb{P}(\mathrm{y}=y|\mathbf{x}=x)}{1-\sum_{j=1}^{J}\mathbb{P}(\mathrm{m}_j=\mathrm{y}|\mathbf{x}=\boldsymbol{x})}. \tag{25}$$

Similarly, w.r.t. $g_{\perp,j}\{\boldsymbol{x}\}$ we have

$$\frac{\partial \mathcal{C}[\psi^J_{\text{SM}}]}{\partial g_{\perp,j}\{\boldsymbol{x}\}} = 0 \implies \frac{\exp\{g_{\perp,j}(\boldsymbol{x})\}}{\sum_{y'\in\mathcal{Y}^\perp}\exp\{g_{y'}(\boldsymbol{x})\}} = \frac{\mathbb{P}(\mathrm{m}_j = \mathrm{y}|\mathbf{x}=\boldsymbol{x})}{1-\sum_{j=1}^J \mathbb{P}(\mathrm{m}_j = \mathrm{y}|\mathbf{x}=\boldsymbol{x})}. \tag{26}$$

The above equations hold true for optimal classifier and the rejector. Thus, if we take the decision as in the main text, we are agreeing with the Bayes solution (considering that denominators are same in both the above conditions).

A.3 Proof of Theorem 3.2: Consistency of ψ^J_{OvA}

The proof follows directly from Verma and Nalisnick (2022). However, for completion we provide the full proof below.

For the surrogate prediction functions $g_1,\ldots,g_K,g_{\perp,1},\ldots,g_{\perp,J}$, and the binary classification surrogate loss $\phi:\{-1,1\}\times\mathbb{R}\to\mathbb{R}_+$, ψ^J_{OvA} takes the following pointwise-form:

$$\begin{aligned}\mathcal{C}[\psi^J_{\text{OvA}}] &= \psi^J_{\text{OvA}}(g_1,\ldots,g_K,g_{\perp,1},\ldots,g_{\perp,J};\boldsymbol{x},y,m_1,\ldots,m_J)\\ &= \phi[g_y(\boldsymbol{x})] + \sum_{y'\in\mathcal{Y},y'\neq y}\phi[-g_{y'}(\boldsymbol{x})] \;+\sum_{j=1}^J \phi[-g_{\perp,j}(\boldsymbol{x})] + \sum_{j=1}^J \mathbb{I}[m_j=y]\left(\phi[g_{\perp,j}(\boldsymbol{x})]-\phi[-g_{\perp,j}(\boldsymbol{x})]\right).\end{aligned}$$

We consider the pointwise *inner* ψ_{OvA}*risk* for some $\mathbf{x}=\boldsymbol{x}$ written as follows:

$$\mathcal{C}[\psi^J_{\text{OvA}}] = \mathbb{E}_{\mathrm{y}|\mathbf{x}=\boldsymbol{x}}\mathbb{E}_{\mathrm{m}|\mathbf{x}=\boldsymbol{x},\mathrm{y}=y}\left[\psi^J_{\text{OvA}}(g_1,\ldots,g_K,g_{\perp,1},\ldots,g_{\perp,J};\boldsymbol{x},y,m_1,\ldots,m_J)\right],$$

We expand both the expectations one-by-one below:

$$\begin{aligned}\mathcal{C}[\psi^J_{\text{OvA}}] = \mathbb{E}_{\mathrm{y}|\mathbf{x}=\boldsymbol{x}}\Bigg[&\phi[g_y(\boldsymbol{x})] + \sum_{y'\in\mathcal{Y},y'\neq y}\phi[-g_{y'}(\boldsymbol{x})] \;+\sum_{j=1}^J \phi[-g_{\perp,j}(\boldsymbol{x})]\\ &+\sum_{j=1}^J\left(\sum_{m_j\in\mathcal{M}}\mathbb{P}(\mathrm{m}_j=m_j|\mathbf{x}=\boldsymbol{x},\mathrm{y}=y)\mathbb{I}\left[m_j=y\right]\left(\phi[g_{\perp,j}(\boldsymbol{x})]-\phi[-g_{\perp,j}(\boldsymbol{x})]\right)\right)\Bigg].\end{aligned}$$

Denote $\mathbb{P}(\mathrm{y}=y|\mathbf{x}=\boldsymbol{x})$ as $\eta_y(\boldsymbol{x})$, then

$$\begin{aligned}\mathcal{C}[\psi^J_{\text{OvA}}] = \mathbb{E}_{\mathrm{y}|\mathbf{x}=\boldsymbol{x}}&\sum_{y\in\mathcal{Y}}\left[\phi[g_y(\boldsymbol{x})] + \sum_{y'\in\mathcal{Y},y'\neq y}\phi[-g_{y'}(\boldsymbol{x})]\right]\\ &+\sum_{j=1}^J\Bigg(\phi[-g_{\perp,j}(\boldsymbol{x})]\\ &+\sum_{y\in\mathcal{Y}}\eta_y(\boldsymbol{x})\Big[\sum_{m_j\in\mathcal{M}}\mathbb{P}(\mathrm{m}_j=m_j|\mathbf{x}=\boldsymbol{x},\mathrm{y}=y)\mathbb{I}\left[m_j=y\right]\left(\phi[g_{\perp,j}(\boldsymbol{x})]-\phi[-g_{\perp,j}(\boldsymbol{x})]\right)\Big]\Bigg)\end{aligned}$$

$$\begin{aligned}\mathcal{C}[\psi^J_{\text{OvA}}] = &\sum_{y\in\mathcal{Y}}\eta_y(\boldsymbol{x})\left[\phi[g_y(\boldsymbol{x})] + \sum_{y'\in\mathcal{Y},y'\neq y}\phi[-g_{y'}(\boldsymbol{x})]\right]\\ &+\sum_{j=1}^J\left(\phi[-g_{\perp,j}(\boldsymbol{x})] + \underbrace{\sum_{y\in\mathcal{Y}}\eta_y(\boldsymbol{x})\sum_{m_j\in\mathcal{M}}\mathbb{P}(\mathrm{m}_j=y|\mathbf{x}=\boldsymbol{x},\mathrm{y}=y)}_{\mathbb{P}(\mathrm{m}_j=\mathrm{y}|\mathbf{x}=\boldsymbol{x})}\left(\phi[g_{\perp,j}(\boldsymbol{x})]-\phi[-g_{\perp,j}(\boldsymbol{x})]\right)\right)\end{aligned}$$

$$\mathcal{C}[\psi^J_{\text{OvA}}] = \sum_{y\in\mathcal{Y}} \eta_y(\boldsymbol{x}) \left[\phi[g_y(\boldsymbol{x})] + \sum_{y'\in\mathcal{Y}, y'\neq y} \phi[-g_{y'}(\boldsymbol{x})]\right] + \sum_{j=1}^{J} \left[\phi[-g_{\perp,j}(\boldsymbol{x})] + \mathbb{P}\left(\mathrm{m}_j = \mathrm{y}|\mathbf{x}=\boldsymbol{x}\right)\left(\phi[g_{\perp,j}(\boldsymbol{x})] - \phi[-g_{\perp,j}(\boldsymbol{x})]\right)\right].$$

Denote $\mathbb{P}(\mathrm{m}_j = \mathrm{y}|\mathbf{x}=\boldsymbol{x})$ as p_{m_j}, then we have

$$\mathcal{C}[\psi^J_{\text{OvA}}] = \sum_{y\in\mathcal{Y}} \eta_y(\boldsymbol{x}) \left[\phi[g_y(\boldsymbol{x})] + \sum_{y'\in\mathcal{Y}, y'\neq y} \phi[-g_{y'}(\boldsymbol{x})]\right] + \sum_{j=1}^{J} \left[\left(1-p_{\mathrm{m}_j}\right)\phi[-g_{\perp,j}(\boldsymbol{x})] + p_{\mathrm{m}_j}\phi[g_{\perp,j}(\boldsymbol{x})]\right]$$

We further simplify the above equation as follows:

$$\mathcal{C}[\psi^J_{\text{OvA}}] = \sum_{y\in\mathcal{Y}} \left[\eta_y(\boldsymbol{x})\cdot\phi[g_y(\boldsymbol{x})] + (1-\eta_y(\boldsymbol{x}))\cdot\phi[-g_y(\boldsymbol{x})]\right] + \sum_{j=1}^{J}\left[\left(1-p_{\mathrm{m}_j}\right)\phi[-g_{\perp,j}(\boldsymbol{x})] + p_{\mathrm{m}_j}\phi[g_{\perp,j}(\boldsymbol{x})]\right].$$

Thus, we conclude from the above expression that we K + J binary classification problems where the pointwise risk (or inner risk) for the i^{th} binary classification problem is given as $\eta_y(\boldsymbol{x})\,\phi(g_y(\boldsymbol{x})) + (1-\eta_y(\boldsymbol{x}))\,\phi(-g_y(\boldsymbol{x}))$ for $i \in [K]$ and $p_{\mathrm{m}_j}(\boldsymbol{x})\,\phi(g_{\perp,j}(\boldsymbol{x})) + \left(1-p_{\mathrm{m}_j}(\boldsymbol{x})\right)\phi(-g_{\perp,j}(x))$ when $i \in [J]\}$. Thus, minimizer of the inner ψ_{OvA} *-risk* can be analyzed in terms of the pointwise minimizer of the *inner* ϕ*-risk* for each of the K + J sub binary classification problems. Denote the minimizer of pointwise *inner* ψ_{OvA}*-risk* as $\boldsymbol{g}^*$, then the above decomposition means g^*_i corresponds to the minimizer of the *inner* ϕ*-risk* for the ith binary classification problem.

We know that the Bayes solution for the binary classification problem is $\operatorname{sign}\left(\eta(\boldsymbol{x}) - \frac{1}{2}\right)$ where $\eta(\boldsymbol{x})$ denotes $p(\mathrm{y}=1|\mathbf{x}=\boldsymbol{x})$. Now when the binary surrogate loss ϕ is a strictly proper composite loss for binary classification, by the property of strictly proper composite losses, we have $\operatorname{sign}(g^*_y(\boldsymbol{x}))$ would agree with the Bayes solution of the Binary classification, i.e. $g^*_y(\boldsymbol{x}) > 0$ if $\eta_y(\boldsymbol{x}) > \frac{1}{2}$. And similarly $g^*_\perp(\boldsymbol{x}) > 0$ if $p_{\mathrm{m}_j}(\boldsymbol{x}) > \frac{1}{2}$. Furthermore, we have the existence of a continuous and increasing inverse link function γ^{-1} for the binary surrogate ϕ with the property that $\gamma^{-1}\left(g^*_y(\boldsymbol{x})\right)$ would converge to $\eta_y(\boldsymbol{x})$. Similarly, $\gamma^{-1}\left(g^*_{\perp,j}(\boldsymbol{x})\right)$ would converge to $p_{\mathrm{m}_j}(\boldsymbol{x})$.

Thus, when the binary surrogate loss ϕ is a strictly proper composite loss, and the classifier and the rejector are defined as in the main text, the minimizer of the pointwise risk $\mathcal{C}\left[\psi^J_{\text{OvA}}\right]$ agree with the Bayes optimal solution. Thus, ψ^J_{OvA} is a calibrated loss function for L2D w.r.t. 0-1 misclassification loss.

A.4 Inconsistency of the Mixture of Experts Formulation (Hemmer et al., 2022)

Proposition A.4. *L_{MoE} is inconsistent for learning to defer.*

The proof works by construction. Specifically, we construct a distribution over $\mathcal{X}\times\mathcal{Y}$ for which the necessary condition for consistency does not hold true.

Consider we have $\mathcal{X} = \{\boldsymbol{x}\}$ and $\mathcal{Y} = \{0,1\}$, i.e. the input space contains the singleton element $\boldsymbol{x}$ with 2 output labels. We define the following distribution $\mathcal{D}$ such that $\mathbb{P}(\boldsymbol{x},0) = \alpha_0$, $\mathbb{P}(\boldsymbol{x},1) = \alpha_1$. For completion, $\alpha_0 + \alpha_1 = 1$. For simplicity, we consider one expert who predicts correctly with perfect confidence, i.e $\mathrm{m} = \mathrm{y}\ \forall y$. The mixture of experts method works by estimating the allocator scores $w_e(\boldsymbol{x})$ (for expert) and $w_c(\boldsymbol{x})$ (for classifier) such that $w_e(\boldsymbol{x}) + w_c(\boldsymbol{x}) = 1$, and the classifier scores $c_0(\boldsymbol{x}), c_1(\boldsymbol{x})$ with $\sum_{i=0}^{1} c_i(\boldsymbol{x}) = 1$. In such a setting, we have

$$\mathbb{E}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}}\left[L_{\text{MoE}}(F,A,\mathbf{x},\mathrm{y},\mathrm{m})\right] = -\alpha_0\left[\log\left(w_e + w_c\cdot c_0\right)\right] - \alpha_1\left[\log\left(w_e + w_c\cdot c_1\right)\right].$$

It is an easy argument to see that the minimum value of the above expression is 0, i.e. $\inf_{A,F}\mathbb{E}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}}\left[L_{\text{MoE}}(F,A,\mathbf{x},\mathrm{y},\mathrm{m})\right] = 0$.

MoE system decides to defer to the expert if $w_e(\boldsymbol{x}) > w_c(\boldsymbol{x})$. For $\delta > 0$, choose $w_e(\boldsymbol{x}) = 0.5 - \delta$ and $w_c(\boldsymbol{x}) = 0.5 + \delta$. Note that $\forall\ \delta > 0$, the system would always decide not to defer to the expert. Also choose $c_0(\boldsymbol{x}) = 1, c_1(\boldsymbol{x}) = 0$. For such an allocator $\bar{A}$ and the classifier $\bar{F}$,

$$\begin{aligned}\mathbb{E}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}}\left[L_{\text{MoE}}\left(\bar{F}, \bar{A}, \mathbf{x}, \mathrm{y}, \mathrm{m}\right)\right] &= -\alpha_1 \cdot \log\left(0.5 - \delta\right) \\ &\leq -\alpha_1 \cdot (0.5 - \delta - 1) = \alpha_1 \cdot (0.5 + \delta)\,,\end{aligned}$$

where the inequality comes from using $\log(x) \leq x - 1$. Next, choose α_1 such that $\alpha_1 = \frac{\delta}{0.5+\delta}$ (why this is true is left as an exercise to the reader). Combining everything, we have shown that

$$\left|\mathbb{E}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}}\left[L_{\text{MoE}}\left(\bar{F}, \bar{A}, \mathbf{x}, \mathrm{y}, \mathrm{m}\right)\right] - \inf_{A,F}\mathbb{E}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}}\left[L_{\text{MoE}}\left(F, A, \mathbf{x}, \mathrm{y}, \mathrm{m}\right)\right]\right| \leq \delta.$$

Thus, our choice of $\bar{A}$ and $\bar{F}$ satisfy Equation 16 for all $\delta > 0$. Since in our construction, we always allow the decision to be made by the classifier which can only predict class label $h(\boldsymbol{x})$ 0, we have $\mathbb{P}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}}\left[h(\boldsymbol{x}) \neq \mathrm{y}\right] = \alpha_1$. And the Bayes optimal rule $h^*(\boldsymbol{x})$ is to always let the expert make the prediction, thus, $\mathbb{P}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}}\left[h^*(\boldsymbol{x}) \neq \mathrm{y}\right] = 0$. Hence, we have

$$\mathbb{P}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}}\left[h(\boldsymbol{x}) \neq \mathrm{y}\right] = \mathbb{P}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}}\left[h^*(\boldsymbol{x}) \neq \mathrm{y}\right] + \eta,$$

where $\eta = \alpha_1$. Thus, $\forall \epsilon < \kappa, \epsilon > 0$, Equation 17 fails to hold true. Hence, we have shown that the optimization of L_{MoE} allows faulty solutions that may not reach the Bayes optimal predictor.

B Choice of Hyperparameters for Regularized Conformal Ensembles

We begin by giving a brief introduction to the procedure of conformal risk control. For detailed exposition, we refer the reader to the original paper (Angelopoulos et al., 2022).

Conformal risk control (Angelopoulos et al., 2022) is a generalized form of conformal prediction which aims to control any bounded monotone loss function $\ell(\cdot)$ in expectation. In our work, we are interested in False Negative Rate (FNR) as a specific loss function which satisfies the monotonicity property as a function of λ (Equation 15). Given access to the calibration data $\{\boldsymbol{x}_n, y_n\}_{n=1}^{\mathrm{N}}$, the goal in conformal risk control is to find $\hat{\lambda}$ so that the following coverage guarantee holds:

$$\mathbb{E}\left[\ell\left(C_{\hat{\lambda}}(\boldsymbol{x}_{n+1})\right)\right] \leq \alpha.$$

Without loss of generality, we consider ℓ to be a non-increasing function of λ and bounded by a constant B. Procedurally, it works by defining $S\left(\boldsymbol{x}_{1:\mathrm{N}}; \lambda\right) = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \ell\left(C_\lambda\left(\boldsymbol{x}_i\right)\right)$. For $\alpha \in (-\infty, B]$, $\hat{\lambda}$ is then defined as

$$\hat{\lambda} = \inf\left\{\lambda : \frac{n}{n+1}S\left(\boldsymbol{x}_{1:\mathrm{N}}; \lambda\right) + \frac{B}{n+1} \leq \alpha\right\}.$$

Assuming exchangeability on $\ell(C_\lambda(\boldsymbol{x}_i))$, this results in the desired coverage guarantees for $C_{\hat{\lambda}}(\boldsymbol{x}_{n+1})$.

In our work, we use a grid of equally spaced 1500 values in $[0, 1]$ to pick $\hat{\lambda}$. We have two hyperparameters κ and β in the regularized conformal ensemble procedure discussed in Section 5. Algorithm 1 below details the procedure to choose κ.

Algorithm 1 Choice of κ

Data: Error rate: α, # Experts: J, Data: $\{\mathbf{s}^i(\boldsymbol{x}), \mathbf{e}^i(\boldsymbol{x})\}_{i=1}^{\mathrm{N}}$ where $\mathbf{s}^i(\boldsymbol{x}) \in [0,1]^{\mathrm{J}}$ and $\mathbf{e}^i(\boldsymbol{x}) \in \{0,1\}^{\mathrm{J}}$
$B \leftarrow \{\cdot\}$
for $i \in [1, N]$ **do**
  for $j \in [1, J]$ **do**
    if $\mathbb{I}\left\{e_j^i(\boldsymbol{x}) == 1\right\}$ **then**
      $B \leftarrow B \cup \{s_j^i(\boldsymbol{x})\}$
    end
  end
  $S \leftarrow$ sort (B) s.t. $u_i, u_j \in S, i \leq j$, then $u_i \geq u_j$
  $\kappa^* \leftarrow 1 - \alpha$ quantile of S
end

We can employ corrections to account for finite sample size N on line 9. Given this procedure to choose κ^*, one may argue that our choice of κ^* can give us meaningful prediction sets by designing a prediction set as:

$$C_2(\boldsymbol{x}) = \{j : s_j(\boldsymbol{x}) \geq \kappa^*\}.$$

However, our next proposition establishes that $C_\lambda(\boldsymbol{x})$ results in prediction sets at most as large as $C_2(\boldsymbol{x})$.

Proposition B.1. *Define the prediction sets* $C_\lambda(\boldsymbol{x}) = \{j : s_j(\boldsymbol{x}) + \beta\,(s_j(\boldsymbol{x}) - \kappa^*) > 1 - \lambda\}$ *and* $C_2(\boldsymbol{x}) = \{j : s_j(\boldsymbol{x}) \geq \kappa^*\}$, *where* κ^* *is defined as above,* $\beta \geq 0$, $0 \leq \lambda \leq 1$, *then it trivially holds that*

$$C_\lambda(\boldsymbol{x}) \subseteq C_2(\boldsymbol{x}).$$

We tune β in a grid-search manner. The grid size for β is $[3.5, 1e^{-3}]$ with steps of 50 samples. We split the total number of deferred samples into two portions: one for tuning hyperparameters β and κ, another for the regular conformal procedure. 30% of the deferred samples are used to tune the ensembling hyperparameters.

C Experimental Setup for Simulated Experts

For the experiment regarding the overall system accuracy described in Section 7.1 we simulate 10 unique experts of increasing ability. Below you can find the description of the experts' configurations for the studied datasets.

C.1 Hate Speech and Galaxy-Zoo

For `Galaxy-Zoo` and `HateSpeech`, we define the following experts using the human annotations available in the datasets and using various perturbations of these predictions:

1. Human expert: we sample predictions from the provided human annotations.
2. Flipping human expert: Expert who flips the given prediction with some probability $\mathtt{p_{flip}}$.
3. Probabilistic expert: Expert who makes use of the annotations with some probability $\mathtt{p_{annotator}}$, or predicts randomly otherwise.

The whole expert configuration is described in Table 1.

Table 1: Hate Speech and Galaxy-Zoo experts configuration.

	Expert configuration	$\mathtt{p_{flip}}$[%]	$\mathtt{p_{annotator}}$[%]
1	Random Expert	-	-
2	Probabilistic Expert	-	10
3	Flipping Human Expert	50	-
4	Probabilistic Expert	-	75
5	Flipping Human Expert	30	-
6	Flipping Human Expert	20	-
7	Probabilistic Expert	-	85
8	Human Expert	-	-
9	Probabilistic Expert	-	50
10	Human Expert	-	-

C.2 HAM10000

The `HAM10000` dataset (Tschandl et al., 2018) is composed of dermatoscopic images corresponding to 7 diagnostic categories in the realm of pigmented lesions. These 7 categories can be further decomposed into benign: melanocytic nevi (`nv`), benign keratinocytic lesions (`bkl`), dermatofisbromas (`df`) and vascular lesions (`vasc`); and malign: melanomas (`mel`), basal cell carcinomas (`bcc`) and actinic keratoses and intraepithelial carcinomas (`akiec`).

In contrast to the Galaxy-Zoo and Hatespeech dataset, for HAM10000 we do not have individual annotators predictions, but just the ground truth label. Further information can be found in the original dataset description (Tschandl et al., 2018). In order to recreate a setup comparable to a real-world scenario, we create different experts configurations:

1. **Random expert**: This expert predicts randomly among all classes.
2. **Dermatologist expert**: These experts will be specialized in a set of categories, and will predict with probability $\texttt{p}_{\texttt{in}}$. Out of that set, they will predict with probability $\texttt{p}_{\texttt{out}}$.
3. **MLPMixer**: We derive `HAM10000`'s expert predictions from the predictions of an 8-layer MLPMixer (Tolstikhin et al., 2021), which has access to additional metadata such as age, gender, and diagnosis type.

As it can be seen in Table 2, we gradually add experts from a random expert to a final expert which simulates an experienced dermatologist. From Kittler et al. (2002) we know that clinical diagnosis of cutaneous melanoma with the unaided eye is only about 60% accuracy, and that dermatologists equipped with dermatoscope can achieve accuracies of $75\% - 84\%$. That is the reason why we chose for the simulated dermatologist experts to have those probabilities $\texttt{p}_{\texttt{in}}$ and $\texttt{p}_{\texttt{out}}$.

Table 2: HAM10000 experts configuration.

	Expert configuration	$\texttt{p}_{\texttt{in}}$ [%]	$\texttt{p}_{\texttt{out}}$ [%]	Diagnostic Category [in]
1	Random Expert	-	-	[nv, bkl, df, vasc, mel, bcc, akiec]
2	Dermatologist for malign	25	15	[mel, bcc, akiec]
3	Dermatologist for benign	25	15	[nv, bkl, df, vasc]
4	Specialized dermatologist in nv	50	15	[nv]
5	Specialized dermatologist in vasc	70	15	[vasc]
6	Specialized dermatologist in mel	75	15	[mel]
7	Dermatologist for benign	75	25	[nv, bkl, df, vasc]
8	MLP Mixer	-	-	[nv, bkl, df, vasc, mel, bcc, akiec]
9	Experienced dermatologist	80	50	[nv, bkl, df, vasc, mel, bcc, akiec]
10	Experienced dermatologist	80	60	[nv, bkl, df, vasc, mel, bcc, akiec]

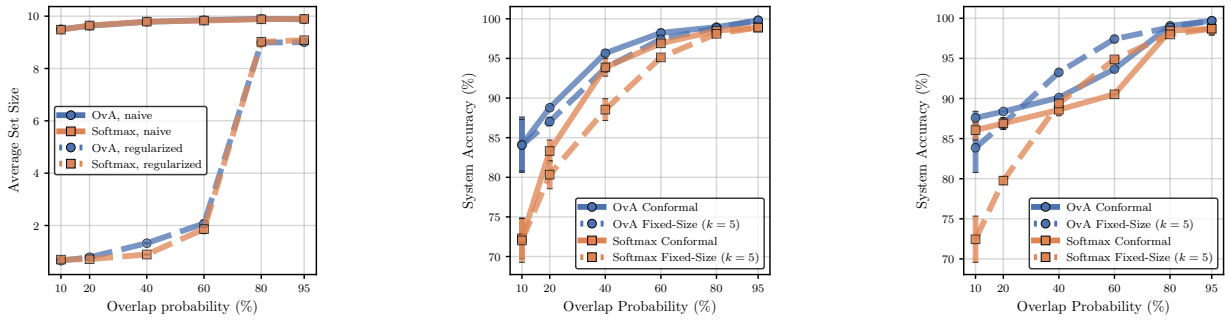
D Additional Experiments and Results

D.1 Overlapping expertise among experts

Similar to the conducted experiments in Section 7.3 in the main manuscript, we want to study ensembling multiple experts, this time under a different experiment setup. We will have 10 experts for the `CIFAR-10` dataset, each of them being an oracle on a specific class out of the 10 classes from the dataset, and we will increase the overlapping probability of these experts being correct on the other classes where the experts are not oracle from 10% to 95% overlap. That is, we will vary from specialized experts to fully overlapped experts. We hope to see that the average set size for specialized experts is close to 1 and for fully overlapped experts close to the total number of experts. We report the results in Figure 4.

Average Set Size First of all, it is worth noticing from Figure 4a that the average set size for the naive conformal method, both for softmax and OvA, is always close to the total number of experts, for specialized and overlapped experts. If we remember, the **naive test statistic** is calculated among *all correct experts*. This is a very important point, because if an expert happens to be correct outside of their expertise domain, this results in a very big non-conformity score because of the low confidence of such expert. That is, imagine for certain sample $\boldsymbol{x}$ and class $y = 3$, for low overlapping probabilities, we might have $E = 3$, where $e = 1$ could be the oracle for class $y = 3$ and $e = 2, e = 3$ two experts that were correct by chance. From Equation 14 in the manuscript, we can expect that, best-case scenario $s_{\pi_1} > s_{\pi_2} > s_{\pi_3}$, or even worse $s_{\pi_1} > \cdots > s_{\pi_8} > s_{\pi_2} > s_{\pi_3}$, because other experts' confidences could be also greater than confidences from correct experts by chance. Therefore, we will obtain bigger test-statistics that result in very larger conformal sets. This problem has already been addressed in Angelopoulos et al. (2022). However, notice how the **regularized test statistic** is capable of producing smaller sets. The idea is that now we optimize additional parameters (described in Section 5 in the manuscript) to ensure that confidences lower than a certain threshold are filtered out for the calculation of the test statistic.

System Accuracies In Figure 4b and 4c we report the system accuracies for the naive conformal method and the regularized conformal method respectively. For the **naive conformal method** we obtain better results than using a fixed-size ensemble of experts of size 5. Because set sizes are almost always close to the total number of experts, and we do majority voting, then as long as the is a correct expert plus an expert correct by chance, we will predict correctly. However, for the **regularized**



(a) Set size. (b) Sys. Acc, naive conformal. (c) Sys. Acc, regularized conformal.

Figure 4: *Gradual overlapping expertise results.* The figures above report the average set size (a) and system accuracies under the naive conformal method (b) and regularized conformal method (c) for increasing expertise overlap for `CIFAR-10`. From Figure (a) we see that the regularized conformal method is more dynamic than the naive method for both OvA and softmax. System accuracies for the naive conformal method (b) are slightly better than fixed-size ensemble because we ensemble all experts, whereas for the regularized conformal method (c) we see a slightly drop due to the adaptivity of the conformal sets

conformal method we notice a drop in the system accuracy for lower overlapping probabilities. Since for such cases now the set sizes are smaller, we have smaller number of experts in the set and therefore less chance of having correct experts by chance in the ensemble. Despite this drop in the accuracy, we clearly have a more dynamic and less conservative ensembling method.