SEPARATING PREDICTION FROM EPISTEMIC UNCERTAINTY IN EVIDENTIAL DEEP LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Evidential deep learning (EDL) predicts a Dirichlet distribution over class-probability vectors in a single forward pass, using its mean for prediction and its dispersion to represent epistemic uncertainty. However, class-label supervision directly informs the predictive mean but provides no direct target for the dispersion. Standard EDL nevertheless uses the same labels to learn both, creating a *supervision–estimand mismatch*: the objective intended to shape dispersion lacks a direct dispersion supervision and can also alter the predictive mean. We address this mismatch with *Bayesian Evidential Dispersion Inference* (BEDI), a probabilistic framework that separates predictive-mean estimation from dispersion inference. BEDI fixes the Dirichlet mean at the calibrated predictive mean of a trained classifier, preserving its class probabilities and predictions by construction. It then infers dispersion using two complementary signals: effective sample size, which quantifies the training information relevant to an input, and Bayesian predictive disagreement, which reflects variation in class probabilities under model uncertainty. Through hierarchical dispersion inference, BEDI uses effective sample size to define a prior over input-specific dispersion and predictive disagreement to update this prior, while learning a shared relationship between effective sample size and dispersion across inputs. This shared relationship enables dispersion inference at new inputs with a single neural-network forward pass. Our theoretical analysis formalizes the supervision–estimand mismatch and establishes contraction of epistemic uncertainty as effective sample size increases. Across standard and long-tailed classification benchmarks and pretrained-model settings, BEDI achieves strong performance on a range of downstream tasks requiring reliable uncertainty quantification.

1 INTRODUCTION

Modern machine learning models achieve remarkable predictive performance, but high-stakes deployment requires more than accuracy: models should also indicate when their predictions may be unreliable. Neural classifiers can be confidently wrong and poorly calibrated (Guo et al., 2017), motivating uncertainty quantification (UQ) to assess the reliability of individual predictions (Gal & Ghahramani, 2016). A conventional classifier outputs a point estimate of class probabilities. While this vector captures predictive uncertainty over the class label, it does not explicitly represent epistemic uncertainty about the estimated class probabilities. Bayesian neural networks and deep ensembles capture this epistemic uncertainty—arising from limited knowledge of the predictor—through variation in class-probability vectors across posterior samples or separately trained ensemble members (Blundell et al., 2015; Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017). However, estimating epistemic uncertainty in this way typically requires multiple forward passes, increasing inference cost and motivating more efficient single-pass alternatives.

Evidential deep learning (EDL) is a prominent single-pass approach that predicts a distribution over class-probability vectors, typically a Dirichlet distribution (Sensoy et al., 2018; Gao et al., 2025). The mean of this distribution determines the predicted class probabilities, while its dispersion is used to represent epistemic uncertainty about those probabilities (Sensoy et al., 2018; Charpentier et al., 2020). By modeling both quantities, EDL can associate the same predicted class probabilities with different levels of epistemic uncertainty. Recent work has refined evidential objectives, relaxed modeling assumptions, and introduced more flexible distributions over class-probability vectors (Deng et al., 2023; Chen et al., 2024; Yoon & Kim, 2024; Chen et al., 2025; Yoon & Kim, 2026a;b).

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

A fundamental challenge, however, is how to learn the dispersion from class-label supervision. While class labels provide information about the mean class probabilities, they do not directly identify the dispersion: distributions over class probabilities with the same mean but different dispersions induce the same class-label probabilities (Bengs et al., 2022; 2023; Juergens et al., 2024). The dispersion learned by EDL must therefore be shaped by the particular training objective beyond what is identified by the class-label likelihood. Indeed, analyses of representative EDL objectives show that the learned dispersion can remain nonzero even in the infinite-data limit, when the class probabilities are fully determined and epistemic uncertainty should vanish (Shen et al., 2024). Thus, explicitly predicting a distribution over class probabilities provides a mechanism for representing epistemic uncertainty, but does not ensure that its learned dispersion faithfully reflects epistemic uncertainty.

This limitation gives rise to two distinct difficulties in standard EDL. First, the same concentration parameters jointly determine the predictive mean and dispersion (Sensoy et al., 2018). As a result, objectives designed to control dispersion can also alter the predictive mean, potentially shifting it away from the true conditional class probabilities and degrading calibration. We call this *predictive-mean interference*. Second, because class-label supervision identifies the predictive mean but not the dispersion, the learned dispersion lacks an identified input-specific target. We call this *dispersion underidentification*. Together, these two difficulties expose a mismatch between what class labels identify—the predictive mean—and what standard EDL jointly estimates—the predictive mean and dispersion. We refer to this discrepancy as the *supervision–estimand mismatch* and formalize it theoretically.

Addressing this mismatch motivates an *estimand-aligned learning principle*: each learning signal should be used only for the estimand that it statistically informs. Following this principle, we pursue two objectives: (i) learning dispersion without altering the predictive mean learned from class labels and (ii) inferring epistemically meaningful dispersion without a directly observed dispersion target. To achieve these objectives, we propose *Bayesian Evidential Dispersion Inference* (BEDI), a probabilistic framework that preserves a calibrated predictive mean while learning higher dispersion where effective training data are scarce or predictive models disagree.

For the first objective, BEDI fixes the mean of the Dirichlet distribution at the calibrated predictive mean of a trained classifier and restricts subsequent inference to dispersion, thereby preserving its calibrated class probabilities and predictions by construction. For the second objective, BEDI uses two signals that capture complementary aspects of epistemic uncertainty. The first is an input-specific effective sample size, which quantifies how much relevant training data are available for a given input; greater effective sample size should correspond to lower dispersion as epistemic uncertainty contracts with increasingly informative data (Bengs et al., 2022; Juergens et al., 2024; Shen et al., 2024). The second is predictive disagreement under an approximate model posterior, which provides an input-specific indication of how uncertain the model remains about the class-probability vector (Gal & Ghahramani, 2016; Shen et al., 2024). Together, these signals characterize how dispersion should vary with both the amount of relevant data and the remaining model uncertainty, without requiring dispersion itself to be observed.

BEDI combines these two signals through hierarchical dispersion inference, giving each a distinct role. Effective sample size determines a prior for dispersion, encoding the expectation that dispersion decreases as more relevant training data become available. Predictive disagreement then provides an input-specific signal for updating this prior. Because disagreement is a model-derived signal rather than an observed dispersion target, BEDI uses it as a noisy pseudo-observation whose reliability determines its influence on the dispersion prior. The model is hierarchical: each input has its own dispersion, while the relationship between effective sample size and dispersion is shared across inputs and learned jointly from their disagreement-derived signals. At test time, BEDI uses this learned relationship to infer dispersion at new inputs with a single neural-network forward pass.

Our theoretical analysis formalizes the supervision–estimand mismatch for the expected mean-squared-error objective of standard EDL and characterizes how effective sample size and predictive disagreement are combined to infer dispersion. We further establish that, with other model quantities fixed, both dispersion and the associated epistemic uncertainty contract as effective sample size increases. Across standard and long-tailed classification benchmarks and pretrained vision, tabular, and language models, BEDI achieves strong performance on a range of downstream UQ tasks.

2 Related Work

Evidential UQ. Recent work has modified both the training and modeling components of standard EDL (Sensoy et al., 2018). $\mathcal{I}$-EDL (Deng et al., 2023) reweights the training loss using Fisher information, whereas R-EDL (Chen et al., 2024) and Re-EDL (Chen et al., 2025) relax fixed choices in the Dirichlet parameterization and training objective. Other approaches have introduced more expressive distributions over class-probability vectors, including $\mathcal{F}$-EDL (Yoon & Kim, 2026b) and MoDEX (Yoon & Kim, 2026a), while DAPPr (Ni et al., 2026) adopts a possibilistic representation of uncertainty. However, standard EDL formulations learn both the predictive mean and dispersion from class-label supervision, although the class-label likelihood identifies only the mean. Prior work has therefore questioned whether label-based objectives can learn dispersion that faithfully reflects epistemic uncertainty (Bengs et al., 2022; 2023; Juergens et al., 2024; Shen et al., 2024). BEDI addresses this issue by separating dispersion inference from predictive-mean learning.

Post-hoc UQ. Post-hoc UQ methods augment a pretrained classifier with an uncertainty model. DMM (Shen et al., 2023) learns a Bayesian Dirichlet meta-model on frozen representations using class-label supervision. STS (Esaki et al., 2024) learns uncertainty from synthetic class-probability targets while preserving the classifier's predicted probabilities. ETN (Chun et al., 2026) uses pretrained logits to learn input-dependent Dirichlet parameters from class-label supervision. BEDI similarly preserves the pretrained classifier, but differs in how dispersion is learned: effective sample size defines its prior structure, while Bayesian predictive disagreement provides input-specific information for updating that prior.

Distance- and density-aware UQ. A complementary line of work grounds uncertainty in feature-space geometry, distance, and density. DUQ (Van Amersfoort et al., 2020) and SNGP (Liu et al., 2020) incorporate feature-space distance into uncertainty estimation. DDU (Mukhoti et al., 2023) estimates uncertainty using feature-space density, while NUQ (Kotelevskii et al., 2022) derives uncertainty from kernel regression in the feature space. Among evidential approaches, PostNet (Charpentier et al., 2020), NatPN (Charpentier et al., 2021), and DAEDL (Yoon & Kim, 2024) use feature-space density to estimate Dirichlet-based uncertainty. In contrast, BEDI uses effective sample size to define a prior over dispersion and Bayesian predictive disagreement to update this prior, so that dispersion reflects both training-data information and model uncertainty rather than feature-space information alone.

Bayesian UQ. Bayesian UQ represents parameter uncertainty through a posterior over model parameters and propagates it to prediction. Since exact inference is generally intractable, methods such as BBB (Blundell et al., 2015) and MC Dropout (Gal & Ghahramani, 2016) use approximate inference, while LLLA (Kristiadi et al., 2020) and VBLL (Harrison et al., 2024) restrict Bayesian inference to the final layer. BEDI uses predictive disagreement from a Bayesian approximation as a training signal for dispersion, enabling single-pass dispersion inference at test time.

3 BEDI: Bayesian Evidential Dispersion Inference

BEDI infers dispersion by combining an effective-sample-size-based prior with Bayesian predictive disagreement while keeping the calibrated predictive mean fixed. We first define a Dirichlet family with a fixed predictive mean, leaving dispersion as the quantity to be inferred (Sec. 3.1). We then construct a prior that relates dispersion to effective sample size (Sec. 3.2.1) and use Bayesian predictive disagreement to update this prior with input-specific information (Sec. 3.2.2). Finally, we develop a hierarchical inference procedure that learns the shared relationship between effective sample size and dispersion across inputs (Sec. 3.3) and uses this relationship to infer dispersion for new inputs at test time (Sec. 3.4).

3.1 Fixed Predictive Mean and Latent Dispersion

We first fix the predictive mean at the calibrated class probabilities of a trained classifier, ensuring that subsequent dispersion inference does not alter its predictions. For a K-class classifier, denote the representation of an input $\mathbf{x}$ by $\mathbf{r}(\mathbf{x})$, and let W and $\mathbf{b}$ denote the weight matrix and bias vector of its classifier head. We apply temperature scaling (Guo et al., 2017) and define

$$\overline{W} = W/T, \qquad \overline{\mathbf{b}} = \mathbf{b}/T, \qquad \widetilde{\mathbf{m}}(\mathbf{x}) \coloneqq \mathrm{softmax}\big(\overline{W}\mathbf{r}(\mathbf{x}) + \overline{\mathbf{b}}\big)\,. \tag{1}$$

The temperature $T > 0$ is fitted by minimizing the negative log-likelihood on a held-out development set $\mathcal{D}_{\mathrm{dev}}$. The feature extractor, classifier head, and temperature are then frozen.

To represent epistemic uncertainty without altering this predictive mean, we place a Dirichlet distribution over the class-probability vector $\boldsymbol{\pi}$, parameterized by a scalar dispersion $d \in (0, 1)$:

$$\boldsymbol{\pi} \mid \mathbf{x}, d \sim \mathrm{Dir}\left((d^{-1} - 1)\widetilde{\mathbf{m}}(\mathbf{x})\right), \qquad \boldsymbol{\pi} \in \Delta^{K-1}. \tag{2}$$

By construction, $\mathbb{E}[\boldsymbol{\pi} \mid \mathbf{x}, d] = \widetilde{\mathbf{m}}(\mathbf{x})$ and $\mathrm{Cov}(\boldsymbol{\pi} \mid \mathbf{x}, d) = dB(\mathbf{x})$, where $B(\mathbf{x}) = \mathrm{Diag}(\widetilde{\mathbf{m}}(\mathbf{x})) - \widetilde{\mathbf{m}}(\mathbf{x})\widetilde{\mathbf{m}}(\mathbf{x})^\top$. Thus, d controls only the dispersion around the fixed predictive mean: smaller d yields greater concentration around the mean, whereas larger d yields greater dispersion. We allow dispersion to vary with the input. Since $d(\mathbf{x}) \in (0, 1)$, we perform inference on its unconstrained logit $z(\mathbf{x}) = \mathrm{logit}(d(\mathbf{x})) \in \mathbb{R}$ with the inverse mapping $d(\mathbf{x}) = \mathrm{sigmoid}(z(\mathbf{x}))$ ensuring that dispersion remains in its valid range. The remaining task is to infer the input-specific latent dispersion $z(\mathbf{x})$ without directly observed dispersion targets.

3.2 Prior and Pseudo-Likelihood for Latent Dispersion

With the predictive mean fixed, we infer dispersion from two complementary sources: effective sample size, which quantifies the amount of relevant training data for an input, and Bayesian predictive disagreement, which captures input-specific variation in class probabilities under an approximate model posterior.

3.2.1 Prior Based on Effective Sample Size

To construct the prior over dispersion, we first build kernel features that capture similarities between a query $\mathbf{x}$ and the training inputs. Using these features in an auxiliary Bayesian linear model, we define the effective sample size at a query as the number of independent observations at that query that would provide the same reduction in output variance as all training inputs. The effective sample size then determines the mean of an input-dependent Gaussian prior over $z(\mathbf{x})$, while the kernel approximation residual adjusts the prior variance.

Nyström kernel approximation for effective sample size. To compute effective sample size scalably, we approximate a kernel on normalized classifier representations using a finite-dimensional Nyström map (Williams & Seeger, 2000; Yang et al., 2012). Let $\mathbf{u}(\mathbf{x}) = \mathbf{r}(\mathbf{x})/\|\mathbf{r}(\mathbf{x})\|_2$ denote the normalized frozen representation. We construct the Nyström map from a smaller set of inducing points obtained by applying k-means separately to each class's normalized training representations, with the same number of centroids per class to prevent frequent classes from dominating the inducing set. Let $\mathcal{Z} = \{\boldsymbol{\zeta}_m\}_{m=1}^M$ denote the resulting inducing set of M centroids. To measure similarity between normalized representations and these centroids, we use an isotropic Gaussian radial basis function (RBF) kernel κ and construct the regularized Nyström map $\boldsymbol{\phi}_{\mathcal{Z}}(\mathbf{u}) = (\mathbf{K}_{\mathcal{Z}\mathcal{Z}} + \lambda_Z I_M)^{-1/2}\boldsymbol{\kappa}_{\mathcal{Z}}(\mathbf{u})$, where $[\boldsymbol{\kappa}_{\mathcal{Z}}(\mathbf{u})]_m = \kappa(\mathbf{u}, \boldsymbol{\zeta}_m)$, $[\mathbf{K}_{\mathcal{Z}\mathcal{Z}}]_{mn} = \kappa(\boldsymbol{\zeta}_m, \boldsymbol{\zeta}_n)$, and $\lambda_Z > 0$ is a small ridge parameter (Appendix A.1).

Effective sample size from variance reduction. We define the effective sample size at a query $\mathbf{x}$ as the number of independent observations at that query that would yield the same reduction in output variance as all training inputs. To construct this, we use an auxiliary Bayesian linear model defined on the regularized Nyström features. Since its posterior covariance depends only on the training inputs, the effective sample size can be computed without observing or generating auxiliary response values.

Let $\{\mathbf{x}_j^{\mathrm{tr}}\}_{j=1}^{N_{\mathrm{tr}}}$ denote the training inputs used to fit the classifier, and let $\Phi_{\mathrm{tr}} = [\boldsymbol{\phi}_{\mathcal{Z}}(\mathbf{u}(\mathbf{x}_j^{\mathrm{tr}}))^\top]_{j=1}^{N_{\mathrm{tr}}}$ denote the corresponding design matrix. We consider the auxiliary linear model $h_{\mathbf{a}}(\mathbf{x}) = \boldsymbol{\phi}_{\mathcal{Z}}(\mathbf{u}(\mathbf{x}))^\top\mathbf{a}$, with $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, I_M)$, and hypothetical observations $y_j^{\mathrm{aux}} = h_{\mathbf{a}}(\mathbf{x}_j^{\mathrm{tr}}) + \epsilon_j$, where $\epsilon_j \sim \mathcal{N}(0, \sigma_n^2)$ and $\sigma_n^2 > 0$. Before conditioning on the training design, the output variance at $\mathbf{x}$ is $v_0(\mathbf{x}) := \mathrm{Var}[h_{\mathbf{a}}(\mathbf{x})] = \|\boldsymbol{\phi}_{\mathcal{Z}}(\mathbf{u}(\mathbf{x}))\|_2^2$. Conditioning on the auxiliary observations yields the posterior covariance $\Sigma_S = (I_M + \sigma_n^{-2}\Phi_{\mathrm{tr}}^\top\Phi_{\mathrm{tr}})^{-1}$ for $\mathbf{a}$, and hence the posterior output variance at the query $v_S(\mathbf{x}) := \boldsymbol{\phi}_{\mathcal{Z}}(\mathbf{u}(\mathbf{x}))^\top\Sigma_S\boldsymbol{\phi}_{\mathcal{Z}}(\mathbf{u}(\mathbf{x}))$. Since the auxiliary response values affect the posterior mean but not Σ_S, $v_S(\mathbf{x})$ depends only on the training inputs.

Now consider n hypothetical, conditionally independent observations of the auxiliary response at the query $\mathbf{x}$, each with observation-noise variance σ_n^2. Under the auxiliary model, these observations would reduce the output variance from $v_0(\mathbf{x})$ to $\left(v_0(\mathbf{x})^{-1} + n\sigma_n^{-2}\right)^{-1}$. We define $n_{\mathrm{eff}}(\mathbf{x})$ as the number of such observations that would yield the same output variance $v_S(\mathbf{x})$ as conditioning on all training inputs. Equating the two variances yields

$$n_{\mathrm{eff}}(\mathbf{x}) := \sigma_n^2\left(\frac{1}{v_S(\mathbf{x})} - \frac{1}{v_0(\mathbf{x})}\right). \tag{3}$$

To obtain a comparable scale across datasets, we normalize $n_{\text{eff}}(\mathbf{x})$ by its training-set average, $\overline{n}_{\text{eff}}^{\text{tr}} := N_{\text{tr}}^{-1} \sum_{j=1}^{N_{\text{tr}}} n_{\text{eff}}(\mathbf{x}_j^{\text{tr}})$, and define the relative effective sample size as $\widetilde{n}_{\text{eff}}(\mathbf{x}) := n_{\text{eff}}(\mathbf{x})/\overline{n}_{\text{eff}}^{\text{tr}}$.

Hierarchical prior for dispersion. Having computed $\widetilde{n}_{\text{eff}}(\mathbf{x})$, we define an input-dependent Gaussian prior over the logit dispersion $z(\mathbf{x})$. Its mean depends on the relative effective sample size $\widetilde{n}_{\text{eff}}(\mathbf{x})$ and the frozen representation $\mathbf{u}(\mathbf{x})$, whereas its variance reflects the quality of the Nyström approximation. For each input $\mathbf{x}_i$ in the held-out development set $\mathcal{D}_{\text{dev}}$, let $\mathbf{u}_i = \mathbf{u}(\mathbf{x}_i)$ denote its normalized frozen representation, $\tilde{n}_i = \widetilde{n}_{\text{eff}}(\mathbf{x}_i)$ its relative effective sample size, and $z_i = z(\mathbf{x}_i)$ its logit dispersion.

The *prior mean* of z_i, denoted by $\mu_i(\boldsymbol{\theta})$, is parameterized as

$$\mu_i(\boldsymbol{\theta}) = \beta_0 + \boldsymbol{\omega}^\top \mathbf{u}_i - \gamma \log(1 + \tilde{n}_i), \tag{4}$$

where β_0 is a global intercept, $\boldsymbol{\omega}$ captures variation across the frozen representation space, and γ controls the effect of relative effective sample size $\tilde{n}_i$. We parameterize $\gamma = \gamma_{\min} + \text{softplus}(\xi)$, where $\gamma_{\min} > 0$ is fixed, ensuring that $\gamma \geq \gamma_{\min} > 0$ and that the prior mean decreases with relative effective sample size. The term $\beta_0 + \boldsymbol{\omega}^\top \mathbf{u}_i$ provides an input-dependent baseline. We collect the shared parameters as $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\omega}, \xi)$ and place the global prior $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}_0, \sigma_0^2 I)$, with fixed $\boldsymbol{\theta}_0$ and σ_0^2. We infer $\boldsymbol{\theta}$ jointly across $\mathcal{D}_{\text{dev}}$.

For the *prior variance*, we account for the quality of the Nyström approximation through the residual $r_K(\mathbf{x}) := \max\left\{0,\, 1 - \|\boldsymbol{\phi}_{\mathcal{Z}}(\mathbf{u}(\mathbf{x}))\|_2^2\right\}$ (Appendix A.1). We set $\tau_i^2 = \tau_0^2 + r_K(\mathbf{x}_i)$, where $\tau_0^2 > 0$ is the base variance. Thus, a larger approximation residual broadens the prior over z_i, reducing the influence of the effective-sample-size-based prior relative to the disagreement-based pseudo-likelihood. The resulting conditional prior is

$$z_i \mid \mathbf{x}_i, \boldsymbol{\theta} \sim \mathcal{N}\big(\mu_i(\boldsymbol{\theta}), \tau_i^2\big)\,. \tag{5}$$

Together, $p(\boldsymbol{\theta})$ and $p(z_i \mid \mathbf{x}_i, \boldsymbol{\theta})$ define the hierarchical prior.

3.2.2 Pseudo-Likelihood Based on Predictive Disagreement

To update the effective-sample-size-based prior, we use Bayesian predictive disagreement obtained from a class-factorized last-layer Laplace approximation (Kristiadi et al., 2020; Daxberger et al., 2021; Immer et al., 2021). We project the resulting predictive-probability covariance onto the fixed-center Dirichlet family, yielding a noisy *pseudo-observation* of latent dispersion whose reliability is determined by the mismatch between the Laplace-derived and fitted Dirichlet covariances.

Predictive-probability covariance via last-layer Laplace. For the temperature-scaled classifier head in Eq. (1), we use a scale-normalized, class-factorized last-layer Laplace approximation centered at the trained class weights, yielding a covariance Σ_k for each class k (Appendix A.2). For each $\mathbf{x}_i \in \mathcal{D}_{\text{dev}}$, let $\mathbf{u}_i = \mathbf{r}(\mathbf{x}_i)/\|\mathbf{r}(\mathbf{x}_i)\|_2$ denote its normalized representation. To reduce sensitivity to representation magnitude, we evaluate the Laplace weight covariance using $\mathbf{u}_i$, yielding the class-wise disagreement variances $\mathbf{u}_i^\top \Sigma_k \mathbf{u}_i$. Propagating these variances through a first-order softmax linearization gives the probability-space disagreement covariance $C_{\text{LA},i} := J_i \,\text{Diag}\big(\mathbf{u}_i^\top \Sigma_1 \mathbf{u}_i, \ldots, \mathbf{u}_i^\top \Sigma_K \mathbf{u}_i\big)\, J_i^\top$, where $J_i = \text{Diag}(\widetilde{\mathbf{m}}(\mathbf{x}_i)) - \widetilde{\mathbf{m}}(\mathbf{x}_i)\widetilde{\mathbf{m}}(\mathbf{x}_i)^\top$ is the softmax Jacobian evaluated at the temperature-scaled logits.

Dispersion projection onto the fixed-center Dirichlet family. To convert the approximate predictive covariance $C_{\text{LA},i}$ into a scalar dispersion signal, we match it to the covariance structure of the fixed-center Dirichlet family. Let $\widetilde{\mathbf{m}}_i = \widetilde{\mathbf{m}}(\mathbf{x}_i)$ and $B_i = \text{Diag}(\widetilde{\mathbf{m}}_i) - \widetilde{\mathbf{m}}_i \widetilde{\mathbf{m}}_i^\top$. The Dirichlet covariance is dB_i, and fixing $\widetilde{\mathbf{m}}_i$ leaves d as the only quantity to estimate in covariance matching. We estimate d using the following variance-normalized projection:

$$\widehat{d}_{\text{LA},i} = \operatorname*{arg\,min}_{d \in [d_{\min}, d_{\max}]} \left\| (B_i^+)^{1/2} (C_{\text{LA},i} - dB_i)(B_i^+)^{1/2} \right\|_F^2 = \text{clip}_{[d_{\min}, d_{\max}]} \left(\frac{\text{tr}(B_i^+ C_{\text{LA},i})}{K - 1} \right), \tag{6}$$

where $0 < d_{\min} < d_{\max} < 1$ and B_i^+ denotes the Moore–Penrose pseudo-inverse of B_i. The normalization prevents high-variance directions under B_i from dominating the fit, while the pseudo-inverse restricts this normalization to the $(K-1)$-dimensional subspace where probability variation occurs, since $B_i \mathbf{1} = 0$ (Appendix A.2). We then transform the resulting dispersion estimate to the logit-scale pseudo-observation $\widehat{z}_{\text{LA},i} = \text{logit}(\widehat{d}_{\text{LA},i})$ and quantify its reliability using the relative covariance mismatch $\rho_{\text{cov},i} := \|C_{\text{LA},i} - \widehat{d}_{\text{LA},i} B_i\|_F^2 / \|C_{\text{LA},i}\|_F^2$.

Pseudo-likelihood for dispersion. We incorporate $\widehat{z}_{\mathrm{LA},i}$ as a noisy pseudo-observation of z_i through the Gaussian pseudo-likelihood

$$\widetilde{p}(\widehat{z}_{\mathrm{LA},i} \mid z_i) = \mathcal{N}\big(\widehat{z}_{\mathrm{LA},i}; z_i, \sigma^2_{\mathrm{LA},i}\big), \qquad \sigma^2_{\mathrm{LA},i} = \sigma^2_{\mathrm{floor}} + \rho_{\mathrm{cov},i}, \tag{7}$$

where $\sigma^2_{\mathrm{floor}} > 0$ is a fixed variance floor. Thus, a larger covariance mismatch increases the pseudo-likelihood variance, giving $\widehat{z}_{\mathrm{LA},i}$ less influence in updating the prior over z_i.

3.3 Hierarchical Dispersion Inference

We now combine the prior and pseudo-likelihood to jointly infer the shared parameters $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\omega}, \xi)$, which govern the prior-mean relationship in Eq. (4), and the input-specific logit dispersions $z_{1:N}$. Let $\{\mathbf{x}_i\}_{i=1}^N$ denote the inputs in $\mathcal{D}_{\mathrm{dev}}$, where $N = |\mathcal{D}_{\mathrm{dev}}|$, and let $\{\widehat{z}_{\mathrm{LA},i}\}_{i=1}^N$ denote the disagreement-derived pseudo-observations of logit dispersion. The resulting generalized posterior is

$$p_G(\boldsymbol{\theta}, z_{1:N} \mid \mathbf{x}_{1:N}, \widehat{z}_{\mathrm{LA},1:N}) \propto \underbrace{p(\boldsymbol{\theta})}_{\text{global prior}} \prod_{i=1}^{N} \underbrace{\mathcal{N}\big(z_i; \mu_i(\boldsymbol{\theta}), \tau_i^2\big)}_{\text{conditional prior}} \underbrace{\mathcal{N}\big(\widehat{z}_{\mathrm{LA},i}; z_i, \sigma^2_{\mathrm{LA},i}\big)}_{\text{pseudo-likelihood}}. \tag{8}$$

Since $\widehat{z}_{\mathrm{LA},i}$ is a model-derived proxy rather than a directly observed value of z_i, we interpret the Gaussian term as a pseudo-likelihood that updates the prior, with more reliable pseudo-observations receiving greater weight. The conditional Gaussian structure simplifies posterior inference: given $\boldsymbol{\theta}$, the posterior of each z_i is available in closed form. We therefore integrate out $z_{1:N}$ and apply variational inference only to the marginal posterior over $\boldsymbol{\theta}$.

Closed-form conditional posterior for z_i. For fixed $\boldsymbol{\theta}$, Gaussian conjugacy between the conditional prior and pseudo-likelihood in Eq. (8) gives $p_G(z_i \mid \widehat{z}_{\mathrm{LA},i}, \mathbf{x}_i, \boldsymbol{\theta}) = \mathcal{N}(z_i; \overline{\mu}_i, \overline{v}_i)$, where $\overline{v}_i = (\tau_i^{-2} + \sigma^{-2}_{\mathrm{LA},i})^{-1}$. The conditional posterior mean is $\overline{\mu}_i = w_i \widehat{z}_{\mathrm{LA},i} + (1 - w_i)\mu_i(\boldsymbol{\theta})$, where $w_i = \tau_i^2/(\tau_i^2 + \sigma^2_{\mathrm{LA},i})$. Thus, the posterior mean balances the disagreement-derived pseudo-observation and the prior mean according to their precisions, as formalized in Theorem 4.2.

Collapsed variational inference for $\boldsymbol{\theta}$. Integrating out the input-specific logit dispersions $z_{1:N}$ gives the marginal pseudo-likelihood $\widetilde{p}(\widehat{z}_{\mathrm{LA},i} \mid \mathbf{x}_i, \boldsymbol{\theta}) = \mathcal{N}(\widehat{z}_{\mathrm{LA},i}; \mu_i(\boldsymbol{\theta}), V_i)$, where $V_i = \tau_i^2 + \sigma^2_{\mathrm{LA},i}$. Because V_i combines uncertainty from both sources, lower reliability of either the prior or the pseudo-observation increases V_i, reducing that input's influence on learning the shared relationship.

The marginal posterior over $\boldsymbol{\theta}$ is not available in closed form, so we approximate it with a diagonal Gaussian $q_\psi(\boldsymbol{\theta})$ and maximize

$$\mathcal{J}_\beta(\psi) = \sum_{i=1}^{N} \mathbb{E}_{q_\psi(\boldsymbol{\theta})} \left[\log \mathcal{N}(\widehat{z}_{\mathrm{LA},i}; \mu_i(\boldsymbol{\theta}), V_i)\right] - \beta\, \mathrm{KL}(q_\psi(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta})). \tag{9}$$

For $\beta = 1$, maximizing this objective minimizes the KL divergence between $q_\psi(\boldsymbol{\theta})$ and the marginal posterior defined above. Combining $q_\psi(\boldsymbol{\theta})$ with the closed-form conditional posteriors of $z_{1:N}$ then gives an approximation to the joint generalized posterior. Other values of $\beta > 0$ target a tempered marginal posterior whose pseudo-likelihood is raised to the power $1/\beta$ (Appendix A.3).

3.4 Test-Time Inference and Uncertainty Quantification

Let $q_{\psi^\star}(\boldsymbol{\theta})$ denote the learned variational posterior over $\boldsymbol{\theta}$ governing the dispersion prior. For a new input, BEDI infers dispersion from its frozen representation, relative effective sample size, and Nyström residual, together with $q_{\psi^\star}(\boldsymbol{\theta})$. The Laplace approximation is used only during learning to construct the disagreement-derived pseudo-observations and is not required at test time.

Posterior prediction. For a test input $\mathbf{x}_*$, we draw $\boldsymbol{\theta} \sim q_{\psi^\star}$ and compute the prior mean $\mu_{\boldsymbol{\theta},*}$ and variance τ_*^2 as in Sec. 3.2.1. We then sample

$$z_* \mid \mathbf{x}_*, \boldsymbol{\theta} \sim \mathcal{N}\big(\mu_{\boldsymbol{\theta},*}, \tau_*^2\big), \quad d_* = \mathrm{sigmoid}(z_*), \quad \boldsymbol{\pi}_* \mid d_*, \mathbf{x}_* \sim \mathrm{Dir}\big((d_*^{-1} - 1)\widetilde{\mathbf{m}}(\mathbf{x}_*)\big). \tag{10}$$

Let $Q_{\mathbf{x}_*}$ denote the resulting marginal distribution of $\boldsymbol{\pi}_*$ after integrating over $\boldsymbol{\theta}$ and z_*. By construction, its mean remains the calibrated probability vector $\widetilde{\mathbf{m}}(\mathbf{x}_*)$, preserving the class prediction $\widehat{y}_* = \arg\max_k [\widetilde{\mathbf{m}}(\mathbf{x}_*)]_k$. Test-time inference requires only a single neural-network forward pass, followed by lightweight kernel computations and Monte Carlo sampling.

Uncertainty measures. We compute total, aleatoric, and epistemic uncertainty using the standard entropy-based decomposition (Depeweg et al., 2018). Under $Y_* \mid \boldsymbol{\pi}_* \sim \mathrm{Cat}(\boldsymbol{\pi}_*)$,

$$\underbrace{H(\widetilde{\mathbf{m}}(\mathbf{x}_*))}_{\mathrm{TU}(\mathbf{x}_*)} = \underbrace{\mathbb{E}_{Q_{\mathbf{x}_*}}[H(\boldsymbol{\pi}_*)]}_{\mathrm{AU}(\mathbf{x}_*)} + \underbrace{I(Y_*; \boldsymbol{\pi}_* \mid \mathbf{x}_*)}_{\mathrm{EU}(\mathbf{x}_*)}, \tag{11}$$

where TU, AU, and EU denote total, aleatoric, and epistemic uncertainty, respectively.

4 Theoretical Analysis

BEDI offers several theoretical results supporting its epistemic uncertainty modeling (see Appendix B for proofs). First, Proposition 4.1 establishes that the expected-MSE objective of standard EDL (Sensoy et al., 2018) couples dispersion control with predictive-mean estimation. For a fixed predictive mean, the population risk increases with dispersion regardless of the true class probabilities. Moreover, even when the predictive mean matches the true class probabilities, the risk can be reduced by shifting it farther from the uniform class-probability vector. This potential distortion motivates estimating the predictive mean and dispersion separately.

Second, Theorem 4.2 characterizes how the effective-sample-size-based prior and predictive disagreement jointly determine the inferred Dirichlet dispersion. The posterior mean dispersion is approximately a reliability-weighted average of the prior-implied and disagreement-derived dispersions, with an explicit error bound that becomes tighter as the posterior variance decreases and the two logit-dispersion estimates become closer.

Finally, Proposition 4.3 shows that for a fixed predictive mean, epistemic uncertainty increases monotonically with Dirichlet dispersion. Building on this result, Theorem 4.4 establishes that epistemic uncertainty decreases as effective sample size increases, with other predictive quantities held fixed. Together, these results show that BEDI produces epistemic uncertainty that contracts as relevant training information increases, addressing prior concerns that EDL may not faithfully represent epistemic uncertainty (Juergens et al., 2024; Shen et al., 2024).

Proposition 4.1 (**Predictive-mean distortion in standard EDL**). *Fix an input* $\mathbf{x}$ *and let* $\mathbf{p} = (p_1, \ldots, p_K) \in \mathrm{int}(\Delta^{K-1})$ *denote the true conditional class-probability vector, where* $p_k = P(Y = k \mid X = \mathbf{x})$, $k = 1, \ldots, K$ *and* $K \geq 2$. *Let* $\mathbf{e}_y \in \{0,1\}^K$ *denote the one-hot vector for class* y. *For a predictive mean* $\mathbf{m} \in \mathrm{int}(\Delta^{K-1})$ *and dispersion* $d \in (0,1)$, *define the standard EDL expected-MSE loss as* $\ell_{\mathrm{MSE}}(\mathbf{e}_y; \mathbf{m}, d) := \mathbb{E}_{\boldsymbol{\pi} \sim \mathrm{Dir}((d^{-1}-1)\mathbf{m})}[\|\mathbf{e}_y - \boldsymbol{\pi}\|_2^2]$, *and the corresponding conditional population risk as* $R_{\mathbf{x}}(\mathbf{m}, d) := \mathbb{E}_{Y|X=\mathbf{x}}[\ell_{\mathrm{MSE}}(\mathbf{e}_Y; \mathbf{m}, d)]$. *Then* $R_{\mathbf{x}}(\mathbf{m}, d) = \|\mathbf{m} - \mathbf{p}\|_2^2 + 1 - \|\mathbf{p}\|_2^2 + d(1 - \|\mathbf{m}\|_2^2)$, *and hence* $\partial R_{\mathbf{x}}(\mathbf{m}, d)/\partial d = 1 - \|\mathbf{m}\|_2^2 > 0$. *Moreover, at* $\mathbf{m} = \mathbf{p}$,

$$\left\langle \nabla_{\mathbf{m}} R_{\mathbf{x}}(\mathbf{p}, d), \mathbf{p} - K^{-1}\mathbf{1} \right\rangle = -2d(\|\mathbf{p}\|_2^2 - K^{-1}) < 0$$

whenever $\mathbf{p} \neq K^{-1}\mathbf{1}$ *and* $d > 0$. *Hence, even when the predictive mean matches the true class probabilities, the expected-MSE objective can be further reduced by moving it farther from the uniform class-probability vector.*

Theorem 4.2 (**Reliability-weighted posterior dispersion**). *Given the conditional posterior of the logit dispersion* z_i *in Sec. 3.3, let* $w_i = \tau_i^2/(\tau_i^2 + \sigma_{\mathrm{LA},i}^2)$ *and* $\overline{v}_i = (\tau_i^{-2} + \sigma_{\mathrm{LA},i}^{-2})^{-1}$. *Define the posterior mean dispersion as* $\overline{d}_i(\boldsymbol{\theta}) := \mathbb{E}[\mathrm{sigmoid}(z_i) \mid \widehat{z}_{\mathrm{LA},i}, \mathbf{x}_i, \boldsymbol{\theta}]$, *the disagreement-derived dispersion* $\widehat{d}_{\mathrm{LA},i}$ *as in Eq. (6), and the prior-implied dispersion as* $d_{\mathrm{prior},i}(\boldsymbol{\theta}) := \mathrm{sigmoid}(\mu_i(\boldsymbol{\theta}))$. *Then the posterior mean dispersion admits the reliability-weighted decomposition* $\overline{d}_i(\boldsymbol{\theta}) = w_i \widehat{d}_{\mathrm{LA},i} + (1 - w_i) d_{\mathrm{prior},i}(\boldsymbol{\theta}) + \varepsilon_i(\boldsymbol{\theta})$, *where the approximation error satisfies*

$$|\varepsilon_i(\boldsymbol{\theta})| \leq \frac{1}{12\sqrt{3}} \left[\overline{v}_i + w_i(1 - w_i) \left(\widehat{z}_{\mathrm{LA},i} - \mu_i(\boldsymbol{\theta}) \right)^2 \right].$$

Proposition 4.3 (**Monotonicity of epistemic uncertainty with Dirichlet dispersion**). *Fix* $K \geq 2$ *and a predictive mean* $\mathbf{m} \in \mathrm{int}(\Delta^{K-1})$. *For each dispersion* $d \in (0,1)$, *let* $\boldsymbol{\pi}_d \sim \mathrm{Dir}((d^{-1} - 1)\mathbf{m})$ *denote the random class-probability vector, and let* $Y_d \mid \boldsymbol{\pi}_d \sim \mathrm{Cat}(\boldsymbol{\pi}_d)$ *denote the corresponding class label. Define* $\mathrm{EU}_{\mathbf{m}}(d) := I(Y_d; \boldsymbol{\pi}_d)$, *the epistemic uncertainty in Eq. (11) with predictive mean* $\mathbf{m}$ *and dispersion* d. *Then,* $\mathrm{EU}_{\mathbf{m}}(d)$ *is strictly increasing in* d.

Theorem 4.4 (**Contraction of epistemic uncertainty with effective sample size**). *Under the predictive model in Eq. (10), fix* $\widetilde{\mathbf{m}}(\mathbf{x})$, $\mathbf{u}(\mathbf{x})$, $r_K(\mathbf{x})$, *and* $q_{\psi^\star}(\boldsymbol{\theta})$, *and assume* $\gamma \geq \gamma_{\min} > 0$ *almost surely. Let* $\mathrm{EU}_{\mathbf{x}}(n)$ *denote the epistemic uncertainty assigned by* BEDI *to* $\mathbf{x}$ *when its relative sample size* $\widetilde{n}_{\mathrm{eff}}(\mathbf{x})$ *is set to* $n \geq 0$. *Then, for any* $0 \leq n_1 < n_2$, $\mathrm{EU}_{\mathbf{x}}(n_2) < \mathrm{EU}_{\mathbf{x}}(n_1)$.

Table 1: **(Downstream UQ performance)** "Miscl. AUPR" denotes misclassification-detection AUPR. In (a), SVHN | C-100 reports OOD AUPR for each dataset. In (b), Shift/OOD AUPR averages ImageNet shifts, Dry Bean held-out classes, and MMLU subjects. Extended results are in Tables 4 to 8. AUROC results are in Tables 9 and 10, and ID calibration results are in Tables 11 and 12.

(a) From-scratch setting

	CIFAR-10			CIFAR-10-LT ($\rho = 0.01$)		
Method	Test Acc.	Miscl. AUPR	OOD AUPR SVHN | C-100	Test Acc.	Miscl. AUPR	OOD AUPR SVHN | C-100
LLLA	91.97 ± 0.2	98.17 ± 0.5	70.79 ± 9.8 | 76.03 ± 2.7	66.46 ± 1.1	76.71 ± 2.3	44.83 ± 3.8 | 60.34 ± 2.2
VBLL	92.50 ± 0.8	98.52 ± 0.2	81.30 ± 13.9 | 80.55 ± 4.2	66.30 ± 1.8	71.90 ± 1.8	37.62 ± 13.8 | 55.36 ± 3.1
SNGP	92.23 ± 0.7	98.90 ± 0.1	87.23 ± 2.1 | 81.53 ± 2.1	66.77 ± 1.1	75.60 ± 1.3	50.64 ± 4.2 | 62.62 ± 0.7
EDL	88.48 ± 0.3	98.74 ± 0.7	82.32 ± 1.2 | 87.13 ± 0.3	42.62 ± 2.7	82.63 ± 1.7	51.99 ± 3.8 | 66.86 ± 0.9
R-EDL	90.09 ± 0.3	98.98 ± 0.1	85.00 ± 1.2 | 87.73 ± 0.3	63.36 ± 1.0	78.34 ± 1.0	48.71 ± 7.1 | 64.20 ± 1.4
$\mathcal{F}$-EDL	91.19 ± 0.2	99.10 ± 0.0	91.20 ± 1.3 | 88.37 ± 0.3	63.73 ± 1.4	85.99 ± 1.7	62.56 ± 2.8 | 70.18 ± 2.0
MoDEX	92.46 ± 0.2	99.18 ± 0.0	91.58 ± 0.4 | 89.28 ± 0.3	71.53 ± 1.1	**91.90** ± 1.2	72.05 ± 3.5 | 76.52 ± 0.9
DMM	87.10 ± 3.2	98.46 ± 0.2	81.59 ± 7.0 | 84.61 ± 2.5	71.52 ± 0.9	90.15 ± 0.6	74.38 ± 2.6 | 75.81 ± 0.6
ETN	90.70 ± 0.0	98.99 ± 0.1	85.60 ± 1.5 | 86.84 ± 0.3	71.21 ± 1.3	82.06 ± 1.5	49.50 ± 10.0 | 58.13 ± 2.2
BEDI	**92.58** ± 0.5	**99.36** ± 0.1	**93.10** ± 1.2 | **90.35** ± 0.6	**72.04** ± 0.9	91.29 ± 0.6	**81.25** ± 1.8 | **77.39** ± 1.0

(b) Pretrained setting

	Vision / ImageNet			Tabular / Dry Bean			Language / OpenBookQA		
Method	Test Acc.	Miscl. AUPR	Shift AUPR Avg.	Test Acc.	Miscl. AUPR	OOD AUPR Avg.	Test Acc.	Miscl. AUPR	OOD AUPR Avg.
LLLA	80.25 ± 0.0	**95.72** ± 0.1	85.25 ± 0.2	95.49 ± 0.2	99.58 ± 0.0	90.23 ± 0.3	86.67 ± 0.7	**97.76** ± 0.6	67.70 ± 3.9
VBLL	80.25 ± 0.0	95.70 ± 0.1	82.35 ± 0.2	95.48 ± 0.2	99.58 ± 0.0	87.59 ± 1.3	86.60 ± 0.5	96.98 ± 0.6	74.74 ± 4.3
$\mathcal{I}$-EDL	80.23 ± 0.1	89.85 ± 0.0	87.56 ± 0.0	95.45 ± 0.1	**99.66** ± 0.0	33.90 ± 0.3	87.13 ± 0.8	96.52 ± 0.6	74.25 ± 7.1
R-EDL	80.21 ± 0.0	89.82 ± 0.0	86.94 ± 0.0	95.32 ± 0.1	99.20 ± 0.1	43.25 ± 0.5	87.67 ± 0.3	94.41 ± 1.0	75.37 ± 3.0
MoDEX	80.13 ± 0.1	89.44 ± 0.1	87.23 ± 0.0	95.06 ± 0.2	98.66 ± 0.2	70.32 ± 3.5	**87.80** ± 0.4	97.48 ± 0.9	37.67 ± 0.4
DMM	70.92 ± 0.1	87.95 ± 0.3	78.50 ± 0.9	95.41 ± 0.1	99.58 ± 0.0	87.97 ± 3.8	86.60 ± 0.2	96.33 ± 0.4	75.91 ± 2.9
ETN	80.24 ± 0.0	92.11 ± 0.0	83.38 ± 0.6	95.49 ± 0.2	99.46 ± 0.1	86.63 ± 1.4	86.53 ± 0.3	94.82 ± 0.7	44.43 ± 7.8
BEDI	**80.26** ± 0.0	95.66 ± 0.0	**88.62** ± 0.2	**95.61** ± 0.2	**99.66** ± 0.0	**91.71** ± 0.4	86.53 ± 0.3	97.61 ± 0.7	**91.78** ± 7.2

5 Experiments

5.1 Experimental Setup

We evaluate downstream UQ performance in (i) a *from-scratch setting* on standard and long-tailed image benchmarks and (ii) a *pretrained setting* across vision, tabular, and language modalities (Sec. 5.2), followed by a controlled experiment testing whether BEDI distinguishes limited relevant training information from predictive ambiguity (Sec. 5.3).

Tasks and metrics. We evaluate (i) ID classification accuracy, (ii) misclassification detection using negative total uncertainty (correct predictions as positives), and (iii) OOD/shift detection using negative epistemic uncertainty (ID samples as positives). Detection performance is measured by area under the precision–recall curve (AUPR) on a 0–100 scale.

Datasets and baselines. In the *from-scratch setting*, ID datasets are CIFAR-10, CIFAR-100, and CIFAR-10-LT (Cao et al., 2019) with imbalance ratios $\rho \in \{0.1, 0.01\}$. For OOD detection, we use SVHN and CIFAR-100 for CIFAR-10 and CIFAR-10-LT, and SVHN and Tiny ImageNet for CIFAR-100. Baselines include evidential, Bayesian, distance- and density-aware, and post-hoc UQ methods. In the *pretrained setting*, we use ImageNet with ImageNet-A (Hendrycks et al., 2021b), ImageNet-Sketch (Wang et al., 2019), and ImageNet-R (Hendrycks et al., 2021a) as distribution shifts for vision; Dry Bean (Koklu & Ozkan, 2020) with one class held out as OOD for tabular data; and OpenBookQA (Mihaylov et al., 2018) with selected MMLU subsets (Hendrycks et al., 2020) as OOD for language. Baselines include post-hoc UQ and fine-tuned EDL methods.

Implementation. From scratch, we use VGG-16 for CIFAR-10 and CIFAR-10-LT and ResNet-18 for CIFAR-100. For pretrained models, we use ResNet-50 with `ImageNet-1K V2` weights, TabPFN-3 (Grinsztajn et al., 2026), and Llama 3.1 8B (Grattafiori et al., 2024) for vision, tabular, and language, respectively. Further details and results are in Appendices D and E.

5.2 Quantitative Evaluation: Downstream UQ Tasks

In the *from-scratch setting*, Table 1(a) shows that BEDI performs strongly on CIFAR-10 and CIFAR-10-LT ($\rho = 0.01$) across downstream UQ tasks. For predictive accuracy and misclassification detection, BEDI preserves the calibrated predictive mean by construction, avoiding the

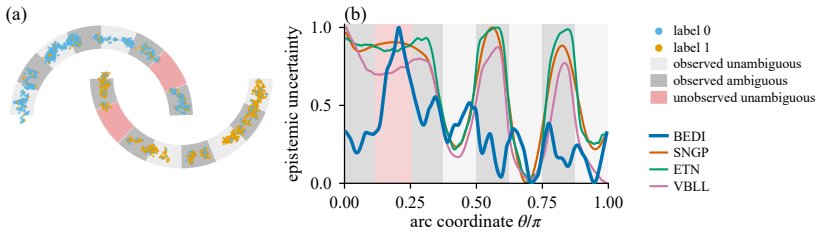

Figure 1: **(Training information vs. predictive ambiguity)** (a) Two-moons benchmark in which training information and predictive ambiguity vary independently. Blue and orange points denote the two classes; light gray, dark gray, and pink backgrounds indicate *observed unambiguous*, *observed ambiguous*, and *unobserved unambiguous* regions, respectively. (b) Min–max-scaled epistemic uncertainty along the arc coordinate θ/π. Shaded bands follow the same color coding as in (a).

predictive-mean distortion characterized in Proposition 4.1. For OOD detection, BEDI achieves strong performance, suggesting that it effectively captures limited relevant training information. This behavior is consistent with its design, which learns from Bayesian predictive disagreement how epistemic uncertainty varies with effective sample size.

In the *pretrained setting*, Table 1(b) shows that BEDI achieves strong downstream UQ performance across vision, tabular, and language models while preserving their predictions. End-to-end UQ methods instead jointly fine-tune the predictive mean and dispersion without direct dispersion supervision, potentially affecting predictive calibration or OOD uncertainty. In contrast, BEDI fixes the predictive mean and infers only dispersion, enabling epistemic uncertainty to be added post hoc to pretrained and foundation models without altering their predictive behavior. Ablations in Table 13 further show that effective sample size and Bayesian predictive disagreement provide complementary information under hierarchical inference.

5.3 Controlled Study: Limited Training Information vs. Predictive Ambiguity

We examine whether epistemic uncertainty distinguishes *limited relevant training information* from *predictive ambiguity*. Epistemic uncertainty should increase as relevant training information becomes limited and contract as informative data accumulate (Juergens et al., 2024; Shen et al., 2024), rather than merely track predictive ambiguity. To test this behavior, we construct a two-moons benchmark in which the availability of relevant training data and label ambiguity vary independently (Fig. 1a). The benchmark includes *observed unambiguous* regions with low label ambiguity, *observed ambiguous* regions with high label ambiguity, and an *unobserved unambiguous* region with low label ambiguity and no training observations.

Several baselines assign low epistemic uncertainty to *observed unambiguous* regions, as expected, but high epistemic uncertainty to *observed ambiguous* regions, making predictive ambiguity difficult to distinguish from limited training information. In contrast, BEDI remains comparatively stable in *observed ambiguous* regions while yielding higher epistemic uncertainty in the *unobserved unambiguous* region (Fig. 1b; Table 14). This behavior reflects BEDI's use of effective sample size to model dispersion, which in turn determines epistemic uncertainty, with Bayesian predictive disagreement informing the learned relationship between effective sample size and dispersion.

6 Conclusion

Summary. We introduced BEDI, a probabilistic framework that separates prediction from dispersion inference for epistemic UQ. BEDI preserves the calibrated predictive mean of a trained classifier while inferring dispersion from effective sample size and Bayesian predictive disagreement within a hierarchical Bayesian framework. Our theoretical analysis motivates this separation and establishes contraction of epistemic uncertainty as relevant training information increases, while experiments across from-scratch and pretrained settings demonstrate strong UQ performance.

Limitations & Future Directions. First, BEDI inherits prediction and representation errors from the frozen classifier, although its post-hoc design readily accommodates stronger predictors. Second, its dispersion estimates depend on the quality of the effective-sample-size and disagreement signals, motivating richer measures of relevant training information and more accurate Bayesian approximations. Third, the current formulation focuses on classification with a Dirichlet distribution over class-probability vectors, leaving extensions to regression and generative modeling for future work.

AI USE STATEMENT

Generative AI tools were used to support manuscript editing for readability and to assist with software code development and debugging. All AI-assisted content was carefully reviewed by the authors, and any code generated with the assistance of generative AI was reviewed and tested for correctness. The authors take full responsibility for the final content of this work, including all text, claims, code, and other materials produced with the assistance of generative AI.

ETHICS STATEMENT

We are not aware of any ethical concerns associated with this work under the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We provide our implementation in the supplementary material and report the experimental configurations and implementation details necessary to reproduce our results in Appendix D.

REFERENCES

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Pitfalls of epistemic uncertainty quantification through loss minimisation. *Advances in Neural Information Processing Systems*, 35: 29205–29216, 2022.

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. On second-order scoring rules for epistemic uncertainty quantification. In *International Conference on Machine Learning*, pp. 2078–2091. PMLR, 2023.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in Neural Information Processing Systems*, 33:1356–1367, 2020.

Bertrand Charpentier, Oliver Borchert, Daniel Zügner, Simon Geisler, and Stephan Günnemann. Natural posterior network: Deep bayesian predictive uncertainty for exponential family distributions. In *International Conference on Learning Representations*, 2021.

Mengyuan Chen, Junyu Gao, and Changsheng Xu. R-edl: Relaxing nonessential settings of evidential deep learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Mengyuan Chen, Junyu Gao, and Changsheng Xu. Revisiting essential and nonessential settings of evidential deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Yongchan Chun, Chanhee Park, Jeongho Yoon, Jaehyung Seo, and Heuiseok Lim. Evidential transformation network: Turning pretrained models into evidential models for post-hoc uncertainty estimation. *arXiv preprint arXiv:2604.08627*, 2026.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in neural information processing systems*, 34:20089–20103, 2021.

Danruo Deng, Guangyong Chen, Yang Yu, Furui Liu, and Pheng-Ann Heng. Uncertainty estimation by fisher information-based evidential deep learning. In *International Conference on Machine Learning*, pp. 7596–7616. PMLR, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, pp. 1184–1193. PMLR, 2018.

Yasushi Esaki, Akihiro Nakamura, Keisuke Kawano, Ryoko Tokuhisa, and Takuro Kutsuna. Accuracy-preserving calibration via statistical modeling on probability simplex. In *International Conference on Artificial Intelligence and Statistics*, pp. 1666–1674. PMLR, 2024.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Junyu Gao, Mengyuan Chen, Liangyu Xiang, and Changsheng Xu. A comprehensive survey on evidential deep learning and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, et al. Tabpfn-3: Technical report. *arXiv preprint arXiv:2605.13986*, 2026.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

James Harrison, John Willes, and Jasper Snoek. Variational bayesian last layers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Sx7BIiPzys.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329. IEEE, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of bayesian neural nets via local linearization. In *International conference on artificial intelligence and statistics*, pp. 703–711. PMLR, 2021.

Mira Juergens, Nis Meinert, Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Is epistemic uncertainty faithfully represented by evidential deep learning methods? In *International Conference on Machine Learning*, pp. 22624–22642. PMLR, 2024.

Murat Koklu and Ilker Ali Ozkan. Multiclass classification of dry beans using computer vision and machine learning techniques. *Computers and Electronics in Agriculture*, 174:105507, 2020.

Nikita Kotelevskii, Aleksandr Artemenkov, Kirill Fedyanin, Fedor Noskov, Alexander Fishkov, Artem Shelmanov, Artem Vazhentsev, Aleksandr Petiushko, and Maxim Panov. Nonparametric uncertainty quantification for single deterministic neural network. *Advances in Neural Information Processing Systems*, 35:36308–36323, 2022.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International conference on machine learning*, pp. 5436–5446. PMLR, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. https://tiny-imagenet.herokuapp.com/, 2015.

Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33:7498–7512, 2020.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2381–2391, 2018.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24384–24394, 2023.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Yao Ni, Jeremie Houssineau, Yew-Soon Ong, and Piotr Koniusz. Possibilistic predictive uncertainty for deep learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=NhL0mBNFk4.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Maohao Shen, Yuheng Bu, Prasanna Sattigeri, Soumya Ghosh, Subhro Das, and Gregory Wornell. Post-hoc uncertainty learning using a dirichlet meta-model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9772–9781, 2023.

Maohao Shen, Jongha Jon Ryu, Soumya Ghosh, Yuheng Bu, Prasanna Sattigeri, Subhro Das, and Gregory Wornell. Are uncertainty quantification capabilities of evidential deep learning a mirage? *Advances in Neural Information Processing Systems*, 37:107830–107864, 2024.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pp. 9690–9700. PMLR, 2020.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019.

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13, 2000.

Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. *Advances in neural information processing systems*, 25, 2012.

Taeseong Yoon and Heeyoung Kim. Uncertainty estimation by density aware evidential deep learning. In *International Conference on Machine Learning*, pp. 57217–57243. PMLR, 2024.

Taeseong Yoon and Heeyoung Kim. Courtroom analogy: New perspective on uncertainty-aware classification. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=dv1mWhkbct.

Taeseong Yoon and Heeyoung Kim. Uncertainty estimation by flexible evidential deep learning. *Advances in Neural Information Processing Systems*, 38:118601–118641, 2026b.

Supplementary Material for

"Separating Prediction from Epistemic Uncertainty in Evidential Deep Learning"

A Derivations of Method Components

- A.1 Nyström Approximation and Effective Sample Size
- A.2 Class-Factorized Laplace Covariance and Whitened Projection
- A.3 Local Gaussian Update and Collapsed Inference
- A.4 Analytic Computation of Predictive Uncertainty

B Proofs of Theoretical Results

- B.1 Proof of Proposition 4.1
- B.2 Proof of Theorem 4.2
- B.3 Proof of Proposition 4.3
- B.4 Proof of Theorem 4.4

C Algorithms

D Experimental Details

- D.1 Datasets
- D.2 Evaluation Protocols
- D.3 Implementation Details
- D.4 Computational Cost and Scalability

E Additional Experimental Results

- E.1 From-Scratch Setting
- E.2 Pretrained Setting
- E.3 Detection AUROC
- E.4 In-Distribution Calibration
- E.5 Ablation Study
- E.6 Controlled Study

A Derivations of Method Components

This section provides derivations for the intermediate results used in the main text.

A.1 Nyström Approximation and Effective Sample Size

This subsection derives the Nyström map, approximation residual, and effective sample size used in Section 3.2.1.

Nyström map and approximation residual. Let $\mathcal{H}_\kappa$ be the reproducing kernel Hilbert space of κ, and write $\boldsymbol{\kappa}_{\mathcal{Z}}(\mathbf{u}) = [\kappa(\mathbf{u}, \boldsymbol{\zeta}_m)]_{m=1}^M$. We begin with the unregularized projection problem

$$E_{\mathbf{u}}(\mathbf{c}) := \left\| \kappa(\mathbf{u}, \cdot) - \sum_{m=1}^{M} c_m \kappa(\boldsymbol{\zeta}_m, \cdot) \right\|_{\mathcal{H}_\kappa}^2 .$$

By the reproducing property, $\nabla_{\mathbf{c}} E_{\mathbf{u}}(\mathbf{c}) = 2\mathbf{K}_{\mathcal{Z}\mathcal{Z}}\mathbf{c} - 2\boldsymbol{\kappa}_{\mathcal{Z}}(\mathbf{u})$. When $\mathbf{K}_{\mathcal{Z}\mathcal{Z}}$ is invertible, the projection coefficients are therefore $\mathbf{c}_0^\star(\mathbf{u}) = \mathbf{K}_{\mathcal{Z}\mathcal{Z}}^{-1}\boldsymbol{\kappa}_{\mathcal{Z}}(\mathbf{u})$. The inner product between the projected kernel sections at $\mathbf{u}$ and $\mathbf{v}$ is

$$\begin{aligned}
\widetilde{\kappa}_{\mathcal{Z}}^0(\mathbf{u}, \mathbf{v}) &= \mathbf{c}_0^\star(\mathbf{u})^\top \mathbf{K}_{\mathcal{Z}\mathcal{Z}} \mathbf{c}_0^\star(\mathbf{v}) \\
&= \boldsymbol{\kappa}_{\mathcal{Z}}(\mathbf{u})^\top \mathbf{K}_{\mathcal{Z}\mathcal{Z}}^{-1} \boldsymbol{\kappa}_{\mathcal{Z}}(\mathbf{v}) \\
&= \left(\mathbf{K}_{\mathcal{Z}\mathcal{Z}}^{-1/2} \boldsymbol{\kappa}_{\mathcal{Z}}(\mathbf{u})\right)^\top \left(\mathbf{K}_{\mathcal{Z}\mathcal{Z}}^{-1/2} \boldsymbol{\kappa}_{\mathcal{Z}}(\mathbf{v})\right).
\end{aligned}$$

Thus, $\mathbf{K}_{\mathcal{Z}\mathcal{Z}}^{-1/2}\boldsymbol{\kappa}_{\mathcal{Z}}(\mathbf{u})$ is a feature map for the unregularized Nyström kernel.

Our method uses ridge regularization. Let $\mathbf{A}_{\mathcal{Z}} = \mathbf{K}_{\mathcal{Z}\mathcal{Z}} + \lambda_Z I_M$, where $\lambda_Z > 0$. Adding a penalty to the projection problem gives

$$\begin{aligned}
J_{\mathbf{u}}(\mathbf{c}) &:= E_{\mathbf{u}}(\mathbf{c}) + \lambda_Z \|\mathbf{c}\|_2^2 \\
&= \kappa(\mathbf{u}, \mathbf{u}) - 2\mathbf{c}^\top \boldsymbol{\kappa}_{\mathcal{Z}}(\mathbf{u}) + \mathbf{c}^\top \mathbf{A}_{\mathcal{Z}} \mathbf{c}.
\end{aligned} \tag{12}$$

Its minimizer is

$$\mathbf{c}^\star(\mathbf{u}) = \mathbf{A}_{\mathcal{Z}}^{-1} \boldsymbol{\kappa}_{\mathcal{Z}}(\mathbf{u}). \tag{13}$$

We define the ridge-regularized Nyström kernel by replacing $\mathbf{K}_{\mathcal{Z}\mathcal{Z}}^{-1}$ in the unregularized kernel above with $\mathbf{A}_{\mathcal{Z}}^{-1}$:

$$\begin{aligned}
\widetilde{\kappa}_{\mathcal{Z}}(\mathbf{u}, \mathbf{v}) &= \boldsymbol{\kappa}_{\mathcal{Z}}(\mathbf{u})^\top \mathbf{A}_{\mathcal{Z}}^{-1} \boldsymbol{\kappa}_{\mathcal{Z}}(\mathbf{v}) \\
&= \boldsymbol{\phi}_{\mathcal{Z}}(\mathbf{u})^\top \boldsymbol{\phi}_{\mathcal{Z}}(\mathbf{v}),
\end{aligned}$$

where $\boldsymbol{\phi}_{\mathcal{Z}}(\mathbf{u}) = \mathbf{A}_{\mathcal{Z}}^{-1/2}\boldsymbol{\kappa}_{\mathcal{Z}}(\mathbf{u})$ is the map used in Section 3.2.1. The approximation coefficients and Nyström features are related by $\mathbf{c}^\star(\mathbf{u}) = \mathbf{A}_{\mathcal{Z}}^{-1/2}\boldsymbol{\phi}_{\mathcal{Z}}(\mathbf{u})$.

Finally, the minimum of the regularized objective gives

$$\begin{aligned}
\min_{\mathbf{c}} J_{\mathbf{u}}(\mathbf{c}) &= \kappa(\mathbf{u}, \mathbf{u}) - \boldsymbol{\kappa}_{\mathcal{Z}}(\mathbf{u})^\top \mathbf{c}^\star(\mathbf{u}) \\
&= \kappa(\mathbf{u}, \mathbf{u}) - \|\boldsymbol{\phi}_{\mathcal{Z}}(\mathbf{u})\|_2^2.
\end{aligned} \tag{14}$$

For the unit-amplitude RBF kernel, $\kappa(\mathbf{u}, \mathbf{u}) = 1$, so this minimum equals the unclipped residual used to define $r_K(\mathbf{x})$ when $\mathbf{u} = \mathbf{u}(\mathbf{x})$. Because $J_{\mathbf{u}}(\mathbf{c}) \geq 0$ and $J_{\mathbf{u}}(\mathbf{0}) = 1$, exact arithmetic gives $0 \leq r_K(\mathbf{x}) \leq 1$. Consequently, $\tau_0^2 \leq \tau_i^2 = \tau_0^2 + r_K(\mathbf{x}_i) \leq \tau_0^2 + 1$. The clipping in r_K only guards against negative roundoff error.

Design-conditioned output variance. Let $\mathbf{y}^{\mathrm{aux}}$ collect the hypothetical responses in the auxiliary Bayesian linear model of Section 3.2.1. With independent Gaussian observation noise of variance $\sigma_n^2 > 0$,

$$\mathbf{y}^{\mathrm{aux}} \mid \mathbf{a}, \Phi_{\mathrm{tr}} \sim \mathcal{N}(\Phi_{\mathrm{tr}}\mathbf{a}, \sigma_n^2 I_{N_{\mathrm{tr}}}).$$

Combining this observation model with the coefficient prior $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, I_M)$ gives

$$\begin{aligned}
-2\log p(\mathbf{a} \mid \mathbf{y}^{\mathrm{aux}}, \Phi_{\mathrm{tr}}) &= \|\mathbf{a}\|_2^2 + \sigma_n^{-2}\|\mathbf{y}^{\mathrm{aux}} - \Phi_{\mathrm{tr}}\mathbf{a}\|_2^2 + \mathrm{const} \\
&= \mathbf{a}^\top (I_M + \sigma_n^{-2}\Phi_{\mathrm{tr}}^\top \Phi_{\mathrm{tr}})\mathbf{a} - 2\sigma_n^{-2}\mathbf{a}^\top \Phi_{\mathrm{tr}}^\top \mathbf{y}^{\mathrm{aux}} + \mathrm{const}.
\end{aligned} \tag{15}$$

The quadratic term determines the posterior covariance, whereas the auxiliary responses occur only in the linear term. Hence,

$$\Sigma_S = (I_M + \sigma_n^{-2}\Phi_{\mathrm{tr}}^\top \Phi_{\mathrm{tr}})^{-1},$$

independently of the response values. Writing $\boldsymbol{\phi}_{\mathbf{x}} = \boldsymbol{\phi}_{\mathcal{Z}}(\mathbf{u}(\mathbf{x}))$, the posterior output variance is $v_S(\mathbf{x}) = \boldsymbol{\phi}_{\mathbf{x}}^\top \Sigma_S \boldsymbol{\phi}_{\mathbf{x}}$, which is the design-conditioned variance defined in Section 3.2.1. Moreover, $0 \prec \Sigma_S \preceq I_M$, so $0 < v_S(\mathbf{x}) \le v_0(\mathbf{x}) = \|\boldsymbol{\phi}_{\mathbf{x}}\|_2^2$ whenever $\boldsymbol{\phi}_{\mathbf{x}} \neq \mathbf{0}$.

Observation-equivalent sample size. Fix $\mathbf{x}$ with $\boldsymbol{\phi}_{\mathbf{x}} \neq \mathbf{0}$ and write $h = h_{\mathbf{a}}(\mathbf{x})$. Before conditioning, $h \sim \mathcal{N}(0, v_0(\mathbf{x}))$. Suppose that n repeated observations at the query satisfy $y_\ell^\star = h + \epsilon_\ell^\star$, with conditionally independent errors $\epsilon_\ell^\star \sim \mathcal{N}(0, \sigma_n^2)$. Their posterior precision for h is $v_0(\mathbf{x})^{-1} + n\sigma_n^{-2}$. Matching the corresponding posterior variance to $v_S(\mathbf{x})$ gives

$$\left(v_0(\mathbf{x})^{-1} + n\sigma_n^{-2}\right)^{-1} = v_S(\mathbf{x}) \quad \Longleftrightarrow \quad n = \sigma_n^2 \left(v_S(\mathbf{x})^{-1} - v_0(\mathbf{x})^{-1}\right). \tag{16}$$

Allowing n to be nonnegative and real-valued yields $n_{\mathrm{eff}}(\mathbf{x})$ in Equation (3). Thus, $n_{\mathrm{eff}}(\mathbf{x})$ is an observation-equivalent count under this repeated-measurement model and need not be an integer. Its nonnegativity follows from $v_S(\mathbf{x}) \le v_0(\mathbf{x})$.

Using the training-set average

$$\overline{n}_{\mathrm{eff}}^{\mathrm{tr}} = \frac{1}{N_{\mathrm{tr}}} \sum_{j=1}^{N_{\mathrm{tr}}} n_{\mathrm{eff}}(\mathbf{x}_j^{\mathrm{tr}}),$$

the relative effective sample size is $\widetilde{n}_{\mathrm{eff}}(\mathbf{x}) = n_{\mathrm{eff}}(\mathbf{x})/\overline{n}_{\mathrm{eff}}^{\mathrm{tr}}$. For the nonempty training design and RBF map used here, each training feature is nonzero and appears in the design, so its output variance is strictly reduced and $\overline{n}_{\mathrm{eff}}^{\mathrm{tr}} > 0$. This normalization therefore gives a mean of one over the training inputs. For each development $\mathbf{x}_i$ in $\mathcal{D}_{\mathrm{dev}}$, the prior uses $\widetilde{n}_i = \widetilde{n}_{\mathrm{eff}}(\mathbf{x}_i)$.

A.2 Class-Factorized Laplace Covariance and Whitened Projection

This subsection derives the class-factorized Laplace covariance and whitened projection used in Section 3.2.2.

Class-factorized Laplace curvature. Let D denote the dimension of $\mathbf{r}(\mathbf{x})$, and let $\overline{\mathbf{w}}_k^\top$ be the kth row of the temperature-scaled weight matrix $\overline{W}$ in Equation (1). For each classifier-training input, write $\mathbf{r}_j^{\mathrm{tr}} = \mathbf{r}(\mathbf{x}_j^{\mathrm{tr}})$ and $\mathbf{u}_j^{\mathrm{tr}} = \mathbf{r}_j^{\mathrm{tr}}/\|\mathbf{r}_j^{\mathrm{tr}}\|_2$, with $\mathbf{r}_j^{\mathrm{tr}} \neq \mathbf{0}$. Let ℓ_j denote the softmax negative log-likelihood for this example under the original classifier head, viewed as a function of the temperature-scaled class weight vectors $\{\mathbf{w}_k\}_{k=1}^K$ with the calibrated bias $\overline{\mathbf{b}}$ held fixed. All derivatives below are evaluated at the trained, temperature-scaled weights $\mathbf{w}_k = \overline{\mathbf{w}}_k$. To reduce sensitivity to representation magnitude, the disagreement construction rescales each data-curvature block by the squared representation norm:

$$\frac{\nabla_{\mathbf{w}_{k'}} \nabla_{\mathbf{w}_k} \ell_j}{\|\mathbf{r}_j^{\mathrm{tr}}\|_2^2} = \widetilde{m}_{jk}\left(\delta_{kk'} - \widetilde{m}_{jk'}\right) \mathbf{u}_j^{\mathrm{tr}} (\mathbf{u}_j^{\mathrm{tr}})^\top, \tag{17}$$

where $\delta_{kk'}$ is the Kronecker delta and $\widetilde{m}_{jk} = [\widetilde{\mathbf{m}}(\mathbf{x}_j^{\mathrm{tr}})]_k$ is evaluated at the original temperature-scaled logits $\overline{W}\mathbf{r}_j^{\mathrm{tr}} + \overline{\mathbf{b}}$. With the Gaussian class-weight prior $p(\mathbf{w}_k) = \mathcal{N}(\mathbf{0}, \lambda_{\mathrm{LA}} I_D)$, where $\lambda_{\mathrm{LA}} > 0$, the scale-normalized curvature blocks are

$$H_{kk'} = \frac{\delta_{kk'}}{\lambda_{\mathrm{LA}}} I_D + \sum_{j=1}^{N_{\mathrm{tr}}} \widetilde{m}_{jk}\left(\delta_{kk'} - \widetilde{m}_{jk'}\right) \mathbf{u}_j^{\mathrm{tr}} (\mathbf{u}_j^{\mathrm{tr}})^\top. \tag{18}$$

The class-factorized approximation discards the blocks with $k \neq k'$ and, when retaining full within-class blocks, sets $\Sigma_k = H_{kk}^{-1}$, so

$$\Sigma_k^{-1} = \frac{1}{\lambda_{\mathrm{LA}}} I_D + \sum_{j=1}^{N_{\mathrm{tr}}} \widetilde{m}_{jk}(1 - \widetilde{m}_{jk}) \mathbf{u}_j^{\mathrm{tr}} (\mathbf{u}_j^{\mathrm{tr}})^\top. \tag{19}$$

For a diagonal implementation, each retained block is additionally diagonalized before inversion, giving instead

$$[\Sigma_k^{-1}]_{dd'} = \delta_{dd'} \left(\lambda_{\mathrm{LA}}^{-1} + \sum_{j=1}^{N_{\mathrm{tr}}} \widetilde{m}_{jk}(1 - \widetilde{m}_{jk})(u_{jd}^{\mathrm{tr}})^2 \right),$$

where $d, d' \in \{1, \dots, D\}$ and u_{jd}^{tr} is the dth coordinate of $\mathbf{u}_j^{\mathrm{tr}}$. In either case, let Σ_k denote the retained covariance. We use the Gaussian approximation $\prod_{k=1}^{K} q_{\mathrm{LA}}(\mathbf{w}_k)$, where $q_{\mathrm{LA}}(\mathbf{w}_k) = \mathcal{N}(\mathbf{w}_k; \overline{\mathbf{w}}_k, \Sigma_k)$. This specifies a frozen-center, scale-normalized Laplace-derived approximation rather than the unmodified posterior curvature of the original raw-feature classifier; the calibrated weights are not reoptimized under this prior. The resulting Σ_k are not, in general, the exact marginal blocks of H^{-1}.

For input $\mathbf{x}_i$ in $\mathcal{D}_{\mathrm{dev}}$, let $\mathbf{r}_i = \mathbf{r}(\mathbf{x}_i)$ and $\mathbf{u}_i = \mathbf{r}_i / \|\mathbf{r}_i\|_2$, with $\mathbf{r}_i \neq \mathbf{0}$. We use the same normalization when evaluating disagreement, with logit perturbation $\delta\ell_{ik} := \mathbf{u}_i^\top \delta\mathbf{w}_k$, where $\delta\mathbf{w}_k = \mathbf{w}_k - \overline{\mathbf{w}}_k$. Independence across classes under the factorized approximation gives

$$\mathrm{Cov}(\delta\boldsymbol{\ell}_i) = \mathrm{Diag}\big(\mathbf{u}_i^\top \Sigma_1 \mathbf{u}_i, \dots, \mathbf{u}_i^\top \Sigma_K \mathbf{u}_i\big)\,. \tag{20}$$

Here, $J_i = \mathrm{Diag}(\widetilde{\mathbf{m}}_i) - \widetilde{\mathbf{m}}_i \widetilde{\mathbf{m}}_i^\top$, with $\widetilde{\mathbf{m}}_i = \widetilde{\mathbf{m}}(\mathbf{x}_i)$, is the softmax Jacobian at the original trained, temperature-scaled logits $\overline{W}\mathbf{r}_i + \overline{\mathbf{b}}$. Linearizing softmax at these unchanged logits under the normalized perturbation gives $\delta\boldsymbol{\pi}_i \approx J_i \delta\boldsymbol{\ell}_i$, and hence

$$\mathrm{Cov}(\delta\boldsymbol{\pi}_i) \approx J_i\, \mathrm{Diag}\big(\mathbf{u}_i^\top \Sigma_1 \mathbf{u}_i, \dots, \mathbf{u}_i^\top \Sigma_K \mathbf{u}_i\big)\, J_i^\top = C_{\mathrm{LA},i}. \tag{21}$$

Thus, $C_{\mathrm{LA},i}$ is the scale-normalized Laplace-derived disagreement covariance used in Section 3.2.2, with normalization applied during both curvature construction and disagreement evaluation.

Variance-normalized projection. Let $\widetilde{\mathbf{m}}_i = \widetilde{\mathbf{m}}(\mathbf{x}_i)$ and $B_i = \mathrm{Diag}(\widetilde{\mathbf{m}}_i) - \widetilde{\mathbf{m}}_i \widetilde{\mathbf{m}}_i^\top$. For an interior calibrated mean, B_i is positive semidefinite with rank $K-1$ and null space $\mathrm{span}\{\mathbf{1}\}$. Since $J_i = B_i$, we also have $C_{\mathrm{LA},i}\mathbf{1} = \mathbf{0}$, so both covariances act on the simplex tangent subspace $\mathbf{1}^\perp$. Write its reduced eigendecomposition as $B_i = U_i \Lambda_i U_i^\top$, where $U_i^\top U_i = I_{K-1}$ and $\Lambda_i \succ 0$. The Moore–Penrose pseudo-inverse is $B_i^+ = U_i \Lambda_i^{-1} U_i^\top$. Define the variance-normalized covariance on this subspace as

$$S_i := \Lambda_i^{-1/2} U_i^\top C_{\mathrm{LA},i} U_i \Lambda_i^{-1/2}. \tag{22}$$

Under the same transformation, dB_i becomes dI_{K-1}. Since U_i has orthonormal columns,

$$\begin{aligned} &\Big\| (B_i^+)^{1/2} (C_{\mathrm{LA},i} - dB_i)(B_i^+)^{1/2} \Big\|_F^2 \\ &\qquad = \|S_i - dI_{K-1}\|_F^2 = \|S_i\|_F^2 - 2d\,\mathrm{tr}(S_i) + (K-1)d^2. \end{aligned} \tag{23}$$

Differentiating with respect to d gives the unconstrained minimizer

$$d_i^\star = \frac{\mathrm{tr}(S_i)}{K-1} = \frac{\mathrm{tr}(B_i^+ C_{\mathrm{LA},i})}{K-1}. \tag{24}$$

Because the objective is a strictly convex quadratic in d, restricting it to $[d_{\min}, d_{\max}]$ clips this minimizer to the interval, yielding $\widehat{d}_{\mathrm{LA},i}$ in Equation (6). The corresponding pseudo-observation is $\widehat{z}_{\mathrm{LA},i} = \mathrm{logit}(\widehat{d}_{\mathrm{LA},i})$. Its pseudo-likelihood variance is determined by

$$\rho_{\mathrm{cov},i} = \frac{\|C_{\mathrm{LA},i} - \widehat{d}_{\mathrm{LA},i} B_i\|_F^2}{\|C_{\mathrm{LA},i}\|_F^2}, \qquad \sigma_{\mathrm{LA},i}^2 = \sigma_{\mathrm{floor}}^2 + \rho_{\mathrm{cov},i}.$$

The denominator is positive for an interior calibrated mean, since $\mathbf{u}_i^\top \Sigma_k \mathbf{u}_i > 0$ for every class. The projection estimates a common covariance scale in the variance-normalized metric, whereas $\rho_{\mathrm{cov},i}$ measures the remaining mismatch in the original probability coordinates.

A.3 Local Gaussian Update and Collapsed Inference

This subsection derives the local Gaussian update and collapsed variational objective used in Section 3.3.

Generalized-Bayes construction. The Gaussian pseudo-likelihood in Section 3.2.2 corresponds to the precision-weighted quadratic loss

$$\mathcal{L}_i(z_i) = \frac{(\widehat{z}_{\mathrm{LA},i} - z_i)^2}{2\sigma_{\mathrm{LA},i}^2}, \qquad \widetilde{p}(\widehat{z}_{\mathrm{LA},i} \mid z_i) = \frac{\exp[-\mathcal{L}_i(z_i)]}{\sqrt{2\pi\sigma_{\mathrm{LA},i}^2}}.$$

Multiplying the hierarchical prior $p(\boldsymbol{\theta}) \prod_{i=1}^{N} \mathcal{N}(z_i; \mu_i(\boldsymbol{\theta}), \tau_i^2)$ by $\exp[-\sum_{i=1}^{N} \mathcal{L}_i(z_i)]$ therefore yields the generalized posterior in Equation (8). Here, $N = |\mathcal{D}_{\text{dev}}|$, and the pseudo-observations and both variance terms are fixed during dispersion inference. This update uses a model-derived discrepancy rather than a likelihood for directly observed dispersion targets.

Local update and marginal factor. For fixed $\boldsymbol{\theta}$, the conditional prior has mean $\mu_i(\boldsymbol{\theta}) = \beta_0 + \boldsymbol{\omega}^\top \mathbf{u}_i - \gamma \log(1 + \widetilde{n}_i)$ and variance τ_i^2. Let $V_i = \tau_i^2 + \sigma_{\text{LA},i}^2$. Completing the square gives

$$\frac{(z_i - \mu_i(\boldsymbol{\theta}))^2}{\tau_i^2} + \frac{(\widehat{z}_{\text{LA},i} - z_i)^2}{\sigma_{\text{LA},i}^2} = \frac{(z_i - \overline{\mu}_i)^2}{\overline{v}_i} + \frac{(\widehat{z}_{\text{LA},i} - \mu_i(\boldsymbol{\theta}))^2}{V_i}, \tag{25}$$

where

$$\overline{v}_i = (\tau_i^{-2} + \sigma_{\text{LA},i}^{-2})^{-1}, \qquad w_i = \frac{\tau_i^2}{\tau_i^2 + \sigma_{\text{LA},i}^2}, \qquad \overline{\mu}_i = w_i \widehat{z}_{\text{LA},i} + (1 - w_i)\mu_i(\boldsymbol{\theta}).$$

Since $\overline{v}_i V_i = \tau_i^2 \sigma_{\text{LA},i}^2$, the normalized Gaussian factors satisfy

$$\begin{aligned} &\mathcal{N}(z_i; \mu_i(\boldsymbol{\theta}), \tau_i^2)\mathcal{N}(\widehat{z}_{\text{LA},i}; z_i, \sigma_{\text{LA},i}^2) \\ &\quad = \mathcal{N}(\widehat{z}_{\text{LA},i}; \mu_i(\boldsymbol{\theta}), V_i)\mathcal{N}(z_i; \overline{\mu}_i, \overline{v}_i). \end{aligned} \tag{26}$$

Normalizing with respect to z_i therefore gives

$$p_G(z_i \mid \widehat{z}_{\text{LA},i}, \mathbf{x}_i, \boldsymbol{\theta}) = \mathcal{N}(z_i; \overline{\mu}_i, \overline{v}_i).$$

Integrating out z_i instead gives the marginal pseudo-likelihood

$$\widetilde{p}(\widehat{z}_{\text{LA},i} \mid \mathbf{x}_i, \boldsymbol{\theta}) = \mathcal{N}(\widehat{z}_{\text{LA},i}; \mu_i(\boldsymbol{\theta}), V_i),$$

which is used in the collapsed objective in Equation (9).

Tempered marginal target. Let $N = |\mathcal{D}_{\text{dev}}|$ and $L_{\text{coll}}(\boldsymbol{\theta}) = \prod_{i=1}^{N} \mathcal{N}(\widehat{z}_{\text{LA},i}; \mu_i(\boldsymbol{\theta}), V_i)$. For $\beta > 0$, define the normalized tempered target

$$p_\beta(\boldsymbol{\theta} \mid \mathbf{x}_{1:N}, \widehat{z}_{\text{LA},1:N}) = \frac{p(\boldsymbol{\theta}) L_{\text{coll}}(\boldsymbol{\theta})^{1/\beta}}{Z_\beta}, \tag{27}$$

where $Z_\beta = \int p(\boldsymbol{\theta}) L_{\text{coll}}(\boldsymbol{\theta})^{1/\beta} \, \mathrm{d}\boldsymbol{\theta}$. Substituting $\log p_\beta = \log p + \beta^{-1} \log L_{\text{coll}} - \log Z_\beta$ into the definition of KL divergence yields

$$\frac{\mathcal{J}_\beta(\psi)}{\beta} = \log Z_\beta - \mathrm{KL}(q_\psi(\boldsymbol{\theta}) \,\|\, p_\beta(\boldsymbol{\theta} \mid \mathbf{x}_{1:N}, \widehat{z}_{\text{LA},1:N})) \,. \tag{28}$$

Thus, maximizing Equation (9) minimizes the KL divergence to p_β within the variational family. For $\beta = 1$, this target is the marginal generalized posterior induced by Equation (8). The associated joint approximation combines the variational global factor with the closed-form conditional factors:

$$q_\psi(\boldsymbol{\theta}, z_{1:N}) = q_\psi(\boldsymbol{\theta}) \prod_{i=1}^{N} \mathcal{N}(z_i; \overline{\mu}_i, \overline{v}_i),$$

where $\overline{\mu}_i$ depends on $\boldsymbol{\theta}$ through $\mu_i(\boldsymbol{\theta})$. For $\beta \neq 1$, tempering affects only the collapsed marginal target; the local Gaussian conditional factors remain unchanged.

A.4 Analytic Computation of Predictive Uncertainty

The aleatoric term in Equation (11) can be evaluated analytically conditional on dispersion. For a test input $\mathbf{x}_*$, write $\mathbf{m} = \widetilde{\mathbf{m}}(\mathbf{x}_*)$ and $s_* = d_*^{-1} - 1 = \exp(-z_*)$. Under the fixed-mean Dirichlet distribution, the power moment is

$$\mathbb{E}[\pi_{*k}^a \mid d_*, \mathbf{x}_*] = \frac{\Gamma(s_*)\Gamma(s_* m_k + a)}{\Gamma(s_* + a)\Gamma(s_* m_k)}, \qquad a > -s_* m_k. \tag{29}$$

Differentiating with respect to a at $a = 1$ yields

$$\mathbb{E}[\pi_{*k} \log \pi_{*k} \mid d_*, \mathbf{x}_*] = m_k \left[\psi_0(s_* m_k + 1) - \psi_0(s_* + 1)\right], \tag{30}$$

where ψ_0 is the digamma function. Using $H(\boldsymbol{\pi}_*) = -\sum_{k=1}^{K} \pi_{*k} \log \pi_{*k}$ and taking the remaining expectation under the posterior predictive construction in Equation (10) gives

$$\begin{aligned} \mathrm{AU}(\mathbf{x}_*) &= \mathbb{E}_{Q_{\mathbf{x}_*}}\left[H(\boldsymbol{\pi}_*)\right] \\ &= \mathbb{E}\left[\psi_0(s_*+1) - \sum_{k=1}^{K} m_k \psi_0(s_* m_k + 1) \,\middle|\, \mathbf{x}_*\right]. \end{aligned} \tag{31}$$

Here, the outer expectation is induced by $\boldsymbol{\theta} \sim q_{\psi^\star}(\boldsymbol{\theta})$, $z_* \mid \mathbf{x}_*, \boldsymbol{\theta} \sim \mathcal{N}(\mu_{\boldsymbol{\theta},*}, \tau_*^2)$, and $d_* = \mathrm{sigmoid}(z_*)$. The integrand is the expected categorical entropy conditional on d_*, not the differential entropy of the Dirichlet distribution. After analytically integrating over $\boldsymbol{\pi}_*$, only an expectation over the scalar dispersion remains. It can be estimated using draws of $\boldsymbol{\theta}$ and z_*, without sampling class-probability vectors. Finally, every conditional Dirichlet distribution has mean $\mathbf{m}$, so the law of total expectation gives

$$\mathbb{E}_{Q_{\mathbf{x}_*}}[\boldsymbol{\pi}_*] = \mathbb{E}_{\boldsymbol{\theta}, z_*}[\mathbb{E}[\boldsymbol{\pi}_* \mid \mathbf{x}_*, \boldsymbol{\theta}, z_*]] = \mathbf{m}.$$

The fixed predictive mean therefore gives

$$\mathrm{TU}(\mathbf{x}_*) = H(\widetilde{\mathbf{m}}(\mathbf{x}_*)), \qquad \mathrm{EU}(\mathbf{x}_*) = \mathrm{TU}(\mathbf{x}_*) - \mathrm{AU}(\mathbf{x}_*).$$

Equivalently, with $Y_* \mid \boldsymbol{\pi}_* \sim \mathrm{Cat}(\boldsymbol{\pi}_*)$,

$$\begin{aligned} \mathrm{EU}(\mathbf{x}_*) &= I(Y_*; \boldsymbol{\pi}_* \mid \mathbf{x}_*) \\ &= \mathbb{E}_{Q_{\mathbf{x}_*}}\left[\mathrm{KL}(\mathrm{Cat}(\boldsymbol{\pi}_*) \,\|\, \mathrm{Cat}(\mathbf{m}))\right]. \end{aligned}$$

This identity follows by expanding the categorical KL divergence and using $\mathbb{E}_{Q_{\mathbf{x}_*}}[\pi_{*k}] = m_k$. Thus, the entropy decomposition applies to the full predictive mixture, without replacing it by a single Dirichlet distribution.

B Proofs of Theoretical Results

B.1 Proof of Proposition 4.1

Proof. Fix an input $\mathbf{x}$ and let $\mathbf{p}$ denote its true conditional class-probability vector, as in the statement. For a fixed class y, the bias–variance decomposition and the Dirichlet moments give

$$\begin{aligned} \ell_{\mathrm{MSE}}(\mathbf{e}_y; \mathbf{m}, d) &= \mathbb{E}_{\boldsymbol{\pi}}\left[\|\mathbf{e}_y - \boldsymbol{\pi}\|_2^2\right] \\ &= \|\mathbf{e}_y - \mathbf{m}\|_2^2 + \mathrm{tr}[\mathrm{Cov}(\boldsymbol{\pi})] \\ &= \|\mathbf{e}_y - \mathbf{m}\|_2^2 + d\left(1 - \|\mathbf{m}\|_2^2\right), \end{aligned} \tag{32}$$

where $\boldsymbol{\pi} \sim \mathrm{Dir}((d^{-1}-1)\mathbf{m})$. Taking expectation over $Y \mid X = \mathbf{x} \sim \mathrm{Cat}(\mathbf{p})$ yields

$$\begin{aligned} R_{\mathbf{x}}(\mathbf{m}, d) &= 1 - 2\mathbf{p}^\top \mathbf{m} + \|\mathbf{m}\|_2^2 + d\left(1 - \|\mathbf{m}\|_2^2\right) \\ &= \|\mathbf{m} - \mathbf{p}\|_2^2 + 1 - \|\mathbf{p}\|_2^2 + d\left(1 - \|\mathbf{m}\|_2^2\right). \end{aligned} \tag{33}$$

Consequently, $\partial R_{\mathbf{x}}(\mathbf{m}, d)/\partial d = 1 - \|\mathbf{m}\|_2^2 > 0$, since $\mathbf{m} \in \mathrm{int}(\Delta^{K-1})$ and $K \geq 2$.

Differentiating with respect to the predictive mean gives

$$\nabla_{\mathbf{m}} R_{\mathbf{x}}(\mathbf{m}, d) = 2(\mathbf{m} - \mathbf{p}) - 2d\mathbf{m}. \tag{34}$$

At $\mathbf{m} = \mathbf{p}$, this gradient is $-2d\mathbf{p}$. The direction $\mathbf{v} = \mathbf{p} - K^{-1}\mathbf{1}$ is tangent to the simplex because $\mathbf{1}^\top \mathbf{v} = 0$. Moreover, since $\mathbf{p}$ is an interior point, $\mathbf{p} + t\mathbf{v}$ remains in the simplex for all sufficiently small $t > 0$. The directional derivative is

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} R_{\mathbf{x}}(\mathbf{p} + t\mathbf{v}, d)\Big|_{t=0} &= \left\langle \nabla_{\mathbf{m}} R_{\mathbf{x}}(\mathbf{p}, d), \mathbf{p} - K^{-1}\mathbf{1} \right\rangle \\ &= -2d\left(\|\mathbf{p}\|_2^2 - K^{-1}\right). \end{aligned} \tag{35}$$

If $\mathbf{p} \neq K^{-1}\mathbf{1}$, then $\|\mathbf{p}\|_2^2 > K^{-1}$, so this derivative is strictly negative for $d > 0$. Consequently, $R_{\mathbf{x}}(\mathbf{p} + t\mathbf{v}, d) < R_{\mathbf{x}}(\mathbf{p}, d)$ for all sufficiently small $t > 0$. Moreover,

$$\begin{aligned} \|\mathbf{p} + t\mathbf{v} - K^{-1}\mathbf{1}\|_2 &= (1+t)\|\mathbf{p} - K^{-1}\mathbf{1}\|_2 \\ &> \|\mathbf{p} - K^{-1}\mathbf{1}\|_2. \end{aligned}$$

Thus, for nonuniform $\mathbf{p}$ and fixed $d > 0$, even when the predictive mean matches the true class probabilities, the expected-MSE objective can be further reduced by moving it farther from the uniform class-probability vector. □

The marginal class-label probability is $\Pr(Y = k \mid \mathbf{m}, d) = \mathbb{E}[\pi_k] = m_k$, independently of d. Thus, the dependence of the expected-MSE objective on dispersion is imposed by the loss rather than identified by the marginal class-label likelihood.

B.2 Proof of Theorem 4.2

Proof. Fix $\boldsymbol{\theta}$. Combining the Gaussian conditional prior and pseudo-likelihood in Equations (5) and (7) and completing the square gives

$$p_G(z_i \mid \widehat{z}_{\mathrm{LA},i}, \mathbf{x}_i, \boldsymbol{\theta}) = \mathcal{N}(z_i; \overline{\mu}_i, \overline{v}_i), \qquad \overline{v}_i = \left(\tau_i^{-2} + \sigma_{\mathrm{LA},i}^{-2}\right)^{-1}, \tag{36}$$

where

$$\begin{aligned} \overline{\mu}_i &= \overline{v}_i \left(\frac{\mu_i(\boldsymbol{\theta})}{\tau_i^2} + \frac{\widehat{z}_{\mathrm{LA},i}}{\sigma_{\mathrm{LA},i}^2} \right) \\ &= w_i \widehat{z}_{\mathrm{LA},i} + (1 - w_i)\mu_i(\boldsymbol{\theta}), \qquad w_i = \frac{\tau_i^2}{\tau_i^2 + \sigma_{\mathrm{LA},i}^2}. \end{aligned} \tag{37}$$

Since $\tau_i^2 > 0$ and $\sigma_{\mathrm{LA},i}^2 > 0$, we have $0 < w_i < 1$ and $\overline{v}_i > 0$. Throughout this proof, expectations over z_i are taken under this conditional generalized posterior.

Let $s(a) = \mathrm{sigmoid}(a)$, so that $\overline{d}_i(\boldsymbol{\theta}) = \mathbb{E}[s(z_i)]$. Its second derivative satisfies

$$s''(a) = s(a)(1 - s(a))(1 - 2s(a)), \qquad \sup_{a \in \mathbb{R}} |s''(a)| = \frac{1}{6\sqrt{3}}. \tag{38}$$

Taylor's theorem at $\mathbb{E}[A]$ and cancellation of the linear term give, for any square-integrable scalar random variable A,

$$|\mathbb{E}[s(A)] - s(\mathbb{E}[A])| \le \frac{\mathrm{Var}(A)}{12\sqrt{3}}. \tag{39}$$

Applying this inequality to the conditional posterior of z_i yields

$$\left|\overline{d}_i(\boldsymbol{\theta}) - s(\overline{\mu}_i)\right| \le \frac{\overline{v}_i}{12\sqrt{3}}. \tag{40}$$

We next compare $s(\overline{\mu}_i)$ with a weighted average on the dispersion scale. For $w \in [0, 1]$, let A equal a with probability w and b with probability $1 - w$. Then $\mathbb{E}[A] = c = wa + (1 - w)b$ and $\mathrm{Var}(A) = w(1 - w)(a - b)^2$. Applying Equation (39) again gives

$$|s(c) - ws(a) - (1 - w)s(b)| \le \frac{w(1 - w)(a - b)^2}{12\sqrt{3}}. \tag{41}$$

Setting $a = \widehat{z}_{\mathrm{LA},i}$, $b = \mu_i(\boldsymbol{\theta})$, and $w = w_i$, so that $c = \overline{\mu}_i$, gives

$$\begin{aligned} &\left|s(\overline{\mu}_i) - w_i s(\widehat{z}_{\mathrm{LA},i}) - (1 - w_i)s(\mu_i(\boldsymbol{\theta}))\right| \\ &\le \frac{w_i(1 - w_i)\left(\widehat{z}_{\mathrm{LA},i} - \mu_i(\boldsymbol{\theta})\right)^2}{12\sqrt{3}}. \end{aligned} \tag{42}$$

By definition, $s(\widehat{z}_{\mathrm{LA},i}) = \widehat{d}_{\mathrm{LA},i}$ and $s(\mu_i(\boldsymbol{\theta})) = d_{\mathrm{prior},i}(\boldsymbol{\theta})$. Adding and subtracting $s(\overline{\mu}_i)$ therefore yields

$$\overline{d}_i(\boldsymbol{\theta}) = w_i \widehat{d}_{\mathrm{LA},i} + (1 - w_i) d_{\mathrm{prior},i}(\boldsymbol{\theta}) + \varepsilon_i(\boldsymbol{\theta}). \tag{43}$$

By Equations (40) and (42) and the triangle inequality, the residual satisfies

$$|\varepsilon_i(\boldsymbol{\theta})| \le \frac{1}{12\sqrt{3}} \left[\overline{v}_i + w_i(1 - w_i)\left(\widehat{z}_{\mathrm{LA},i} - \mu_i(\boldsymbol{\theta})\right)^2\right], \tag{44}$$

which establishes the claimed reliability-weighted decomposition of $\overline{d}_i(\boldsymbol{\theta})$ and its approximation error bound. □

B.3 Proof of Proposition 4.3

Proof. Fix $K \geq 2$ and $\mathbf{m} \in \operatorname{int}(\Delta^{K-1})$. Write $s = d^{-1} - 1 > 0$, so that $d(s) = (1+s)^{-1}$. Since $\mathbb{E}[\boldsymbol{\pi}_d] = \mathbf{m}$, the marginal distribution of Y_d is $\operatorname{Cat}(\mathbf{m})$. Hence,

$$\mathrm{EU}_{\mathbf{m}}(d) = I(Y_d; \boldsymbol{\pi}_d) = H(\mathbf{m}) - \mathbb{E}[H(\boldsymbol{\pi}_d)].$$

For $\boldsymbol{\pi}_d \sim \operatorname{Dir}(s\mathbf{m})$, the Dirichlet power moment is

$$\mathbb{E}[\pi_{d,k}^a] = \frac{\Gamma(s)\Gamma(sm_k + a)}{\Gamma(s+a)\Gamma(sm_k)}, \qquad a > -sm_k.$$

Differentiating with respect to a at $a = 1$ gives

$$\mathbb{E}[\pi_{d,k} \log \pi_{d,k}] = m_k \left[\psi_0(sm_k + 1) - \psi_0(s+1)\right], \tag{45}$$

where ψ_0 is the digamma function. Consequently,

$$\mathrm{EU}_{\mathbf{m}}(d) = \sum_{k=1}^{K} m_k \left[\psi_0(sm_k + 1) - \psi_0(s+1) - \log m_k\right]. \tag{46}$$

Differentiating with respect to s yields

$$\frac{\mathrm{d}}{\mathrm{d}s} \mathrm{EU}_{\mathbf{m}}(d(s)) = \sum_{k=1}^{K} m_k^2 \psi_1(sm_k + 1) - \psi_1(s+1), \tag{47}$$

where ψ_1 is the trigamma function. Its integral representation is

$$\psi_1(a+1) = \int_0^\infty \frac{ue^{-au}}{e^u - 1}\,\mathrm{d}u, \qquad a > 0. \tag{48}$$

Substituting this representation and changing variables in each class-specific term gives

$$\frac{\mathrm{d}}{\mathrm{d}s} \mathrm{EU}_{\mathbf{m}}(d(s)) = \int_0^\infty ue^{-su} \left[\sum_{k=1}^{K} \frac{1}{e^{u/m_k} - 1} - \frac{1}{e^u - 1}\right] \mathrm{d}u. \tag{49}$$

For fixed $u > 0$, define $h_u(a) = (e^{u/a} - 1)^{-1}$ for $a > 0$. The ratio

$$\frac{h_u(a)}{a} = \frac{1}{a(e^{u/a} - 1)} \tag{50}$$

is strictly increasing in a, because its positive denominator satisfies

$$\frac{\mathrm{d}}{\mathrm{d}a}\left[a(e^{u/a} - 1)\right] = e^{u/a}\left(1 - \frac{u}{a}\right) - 1 < 0. \tag{51}$$

Here we used $e^v(1-v) < 1$ for every $v > 0$, which follows because $e^v(1-v)$ equals one at $v = 0$ and has derivative $-ve^v < 0$ for $v > 0$. Since $0 < m_k < 1$, we have $h_u(m_k) < m_k h_u(1)$. Summing over k and using $\sum_k m_k = 1$ gives

$$\sum_{k=1}^{K} \frac{1}{e^{u/m_k} - 1} < \frac{1}{e^u - 1}. \tag{52}$$

Thus, the integrand in Equation (49) is strictly negative for every $u > 0$, and $\mathrm{d}\,\mathrm{EU}_{\mathbf{m}}(d(s))/\mathrm{d}s < 0$. Finally, $\mathrm{d}s/\mathrm{d}d = -d^{-2} < 0$, so

$$\mathrm{EU}'_{\mathbf{m}}(d) = -\frac{1}{d^2}\,\frac{\mathrm{d}}{\mathrm{d}s}\,\mathrm{EU}_{\mathbf{m}}\big((1+s)^{-1}\big)\Big|_{s=d^{-1}-1} > 0. \tag{53}$$

Therefore, $\mathrm{EU}_{\mathbf{m}}(d)$ is strictly increasing on $(0, 1)$. □

B.4 Proof of Theorem 4.4

Proof. Fix $\widetilde{\mathbf{m}}(\mathbf{x})$, $\mathbf{u}(\mathbf{x})$, $r_K(\mathbf{x})$, and $q_{\psi^\star}(\boldsymbol{\theta})$ as in the statement. The fixed-mean Dirichlet construction has $K \geq 2$ and $\widetilde{\mathbf{m}}(\mathbf{x}) \in \mathrm{int}(\Delta^{K-1})$. Let $\tau(\mathbf{x}) = \sqrt{\tau_0^2 + r_K(\mathbf{x})}$. Only the relative effective sample size n is varied below; the fitted quantities and the effective-sample-size normalization are held fixed. On a common probability space, independently draw $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\omega}, \xi) \sim q_{\psi^\star}(\boldsymbol{\theta})$ and $\varepsilon \sim \mathcal{N}(0,1)$. For each $n \geq 0$, define

$$\begin{aligned} Z_{\mathbf{x}}(n) &= \beta_0 + \boldsymbol{\omega}^\top \mathbf{u}(\mathbf{x}) - \gamma \log(1+n) + \tau(\mathbf{x})\varepsilon, \\ D_{\mathbf{x}}(n) &= \mathrm{sigmoid}(Z_{\mathbf{x}}(n)), \end{aligned} \tag{54}$$

where $\gamma = \gamma_{\min} + \mathrm{softplus}(\xi)$. These are coupled versions of the logit dispersion and dispersion under Equation (10) when the relative effective sample size $\widetilde{n}_{\mathrm{eff}}(\mathbf{x})$ is set to n.

For $0 \leq n_1 < n_2$, the same draws of $\boldsymbol{\theta}$ and ε give

$$Z_{\mathbf{x}}(n_2) - Z_{\mathbf{x}}(n_1) = -\gamma \log\left(\frac{1+n_2}{1+n_1}\right) < 0 \qquad \text{almost surely,} \tag{55}$$

since $\gamma \geq \gamma_{\min} > 0$ almost surely. The sigmoid is strictly increasing, so $D_{\mathbf{x}}(n_2) < D_{\mathbf{x}}(n_1)$ almost surely. In particular, since both variables lie in $(0,1)$,

$$\mathbb{E}[D_{\mathbf{x}}(n_2)] < \mathbb{E}[D_{\mathbf{x}}(n_1)]. \tag{56}$$

To relate this ordering to the epistemic uncertainty assigned by BEDI, let $Q_{\mathbf{x}}^{(n)}$ denote the marginal distribution induced by

$$\boldsymbol{\pi} \mid D_{\mathbf{x}}(n) \sim \mathrm{Dir}\big(\big(D_{\mathbf{x}}(n)^{-1} - 1\big)\widetilde{\mathbf{m}}(\mathbf{x})\big)\,.$$

Every conditional Dirichlet distribution has the same mean $\widetilde{\mathbf{m}}(\mathbf{x})$, independently of $D_{\mathbf{x}}(n)$. Thus, the marginal mixture $Q_{\mathbf{x}}^{(n)}$ also has mean $\widetilde{\mathbf{m}}(\mathbf{x})$. Using Equation (11) and the law of total expectation, we therefore obtain

$$\begin{aligned} \mathrm{EU}_{\mathbf{x}}(n) &= H(\widetilde{\mathbf{m}}(\mathbf{x})) - \mathbb{E}_{Q_{\mathbf{x}}^{(n)}}[H(\boldsymbol{\pi})] \\ &= \mathbb{E}[H(\widetilde{\mathbf{m}}(\mathbf{x})) - \mathbb{E}[H(\boldsymbol{\pi}) \mid D_{\mathbf{x}}(n)]] \\ &= \mathbb{E}\big[\mathrm{EU}_{\widetilde{\mathbf{m}}(\mathbf{x})}\big(D_{\mathbf{x}}(n)\big)\big]\,. \end{aligned} \tag{57}$$

By Proposition 4.3, $\mathrm{EU}_{\widetilde{\mathbf{m}}(\mathbf{x})}(d)$ is strictly increasing in d. Hence, under the coupling above,

$$\mathrm{EU}_{\widetilde{\mathbf{m}}(\mathbf{x})}\big(D_{\mathbf{x}}(n_2)\big) < \mathrm{EU}_{\widetilde{\mathbf{m}}(\mathbf{x})}\big(D_{\mathbf{x}}(n_1)\big) \qquad \text{almost surely.} \tag{58}$$

These quantities are bounded between zero and $H(\widetilde{\mathbf{m}}(\mathbf{x})) \leq \log K$. Taking expectations and applying Equation (57) therefore gives

$$\mathrm{EU}_{\mathbf{x}}(n_2) < \mathrm{EU}_{\mathbf{x}}(n_1). \tag{59}$$

Thus, with the quantities specified in the statement held fixed, the epistemic uncertainty assigned by BEDI to $\mathbf{x}$ is strictly decreasing in its relative effective sample size. □

C ALGORITHMS

Algorithms 1 and 2 summarize post-hoc fitting and prediction for BEDI. During fitting, Laplace-derived pseudo-observations are used to learn a variational posterior over dispersion parameters. The inducing points, Nyström kernel factors, auxiliary covariance, and training-set normalization are saved as a support state $\mathcal{S}$ for processing new inputs. Prediction uses this state and the learned posterior without recomputing Laplace quantities.

Algorithm 1 Post-hoc fitting of BEDI

Require: Training set $\mathcal{D}_{\text{train}}$, development set $\mathcal{D}_{\text{dev}}$, trained classifier $(\mathbf{r}, W, \mathbf{b})$, and hyperparameters
Ensure: Frozen calibrated predictor, support state $\mathcal{S}$, and variational posterior $q_{\psi^\star}(\boldsymbol{\theta})$

Calibration and effective sample size

Fit temperature T on $\mathcal{D}_{\text{dev}}$ and freeze the calibrated predictor defined in Equation (1).
Construct the Nyström map $\boldsymbol{\phi}_{\mathcal{Z}}$ using class-balanced inducing points from normalized training representations.
Compute Σ_S and the training-set normalization $\overline{n}_{\text{eff}}^{\text{tr}}$ as in Section 3.2.1.

Disagreement-derived pseudo-observations

Select up to N_{LA} training representations for Laplace construction.
Construct the class-factorized Laplace approximation centered at the temperature-scaled classifier weights, holding the calibrated bias fixed, and obtain $\{\Sigma_k\}_{k=1}^K$ from the scale-normalized curvature in Section A.2.
for each input $\mathbf{x}_i$ in $\mathcal{D}_{\text{dev}}$ **do**
    Compute $\widetilde{\mathbf{m}}_i = \widetilde{\mathbf{m}}(\mathbf{x}_i)$ and $\mathbf{u}_i = \mathbf{r}(\mathbf{x}_i)/\|\mathbf{r}(\mathbf{x}_i)\|_2$.
    Compute relative effective sample size $\widetilde{n}_i$ and prior variance $\tau_i^2 = \tau_0^2 + r_K(\mathbf{x}_i)$ as in Section 3.2.1.
    Compute $C_{\text{LA},i}$ using $\{\mathbf{u}_i^\top \Sigma_k \mathbf{u}_i\}_{k=1}^K$ and construct $(\widehat{z}_{\text{LA},i}, \sigma_{\text{LA},i}^2)$ as in Section 3.2.2.
end for

Posterior learning

Fit the diagonal-Gaussian variational posterior $q_{\psi^\star}(\boldsymbol{\theta})$ by maximizing Equation (9), where $V_i = \tau_i^2 + \sigma_{\text{LA},i}^2$.
Store the inducing points, Nyström kernel factors, Σ_S, $\overline{n}_{\text{eff}}^{\text{tr}}$, and fixed kernel and prior settings as $\mathcal{S}$; discard the Laplace covariances and disagreement-derived pseudo-observations.

Algorithm 2 Posterior prediction with BEDI

Require: Test input $\mathbf{x}_*$, frozen calibrated predictor, support state $\mathcal{S}$, posterior $q_{\psi^\star}(\boldsymbol{\theta})$, and number of draws J
Ensure: Class prediction $\widehat{y}_*$ and uncertainty estimates $\text{TU}(\mathbf{x}_*)$, $\text{AU}(\mathbf{x}_*)$, and $\text{EU}(\mathbf{x}_*)$

Obtain $\widetilde{\mathbf{m}}_* = \widetilde{\mathbf{m}}(\mathbf{x}_*)$ and $\mathbf{u}_* = \mathbf{r}(\mathbf{x}_*)/\|\mathbf{r}(\mathbf{x}_*)\|_2$ with one forward pass through the frozen predictor.
Compute $\widetilde{n}_{\text{eff}}(\mathbf{x}_*)$, $r_K(\mathbf{x}_*)$, and $\tau_*^2 = \tau_0^2 + r_K(\mathbf{x}_*)$ using $\mathcal{S}$ as in Section 3.2.1.
for $j = 1, \ldots, J$ **do**
    Draw $\boldsymbol{\theta}^{(j)} \sim q_{\psi^\star}(\boldsymbol{\theta})$ and compute $\mu_{\boldsymbol{\theta}^{(j)},*}$ using Equation (4).
    Draw $z_*^{(j)} \sim \mathcal{N}(\mu_{\boldsymbol{\theta}^{(j)},*}, \tau_*^2)$ and set $d_*^{(j)} = \text{sigmoid}(z_*^{(j)})$.
    Compute the expected categorical entropy conditional on $d_*^{(j)}$ using Equation (31).
end for
Set $\widehat{y}_* = \arg\max_k [\widetilde{\mathbf{m}}_*]_k$ and $\text{TU}(\mathbf{x}_*) = H(\widetilde{\mathbf{m}}_*)$.
Estimate AU by averaging the conditional entropy terms and set $\text{EU}(\mathbf{x}_*) = \text{TU}(\mathbf{x}_*) - \text{AU}(\mathbf{x}_*)$.

D EXPERIMENTAL DETAILS

This section provides additional details on the experiments reported in the main text. We describe the datasets in Section D.1, the evaluation protocols in Section D.2, and the implementation of BEDI in Section D.3. We analyze computational complexity and report post-hoc fitting and inference costs in Section D.4.

D.1 DATASETS

Our experiments cover two complementary settings. The trained-from-scratch setting uses CIFAR-10, CIFAR-100, and CIFAR-10-LT to evaluate uncertainty quantification under balanced and long-tailed image classification. The pretrained setting considers ImageNet, Dry Bean, and OpenBookQA. The corresponding OOD and distribution-shift benchmarks are described alongside the ID datasets below.

D.1.1 FROM-SCRATCH SETTING

CIFAR-10. This dataset (Krizhevsky et al., 2009) contains color images from ten animal and object categories: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. It consists of 50,000 training images and 10,000 test images, each with resolution 32×32. Both splits are class-balanced, with 5,000 training and 1,000 test images per class. We use it as an ID benchmark for standard image classification and as the source dataset for constructing CIFAR-10-LT. Its OOD evaluations use the test sets of SVHN and CIFAR-100.

CIFAR-100. This dataset (Krizhevsky et al., 2009) contains 100 animal and object categories at the same 32×32 resolution as CIFAR-10. It has 50,000 training images and 10,000 test images, corresponding to 500 training and 100 test images per class. Relative to CIFAR-10, it evaluates uncertainty estimation with a larger label space and fewer training examples per class. We use it as an ID benchmark with SVHN and Tiny ImageNet-200 as OOD datasets. Its official test set also serves as OOD data for the CIFAR-10 and CIFAR-10-LT experiments.

CIFAR-10-LT. This dataset is a long-tailed variant of CIFAR-10 constructed by exponentially subsampling its training classes (Cao et al., 2019). After ordering classes from largest to smallest, the retained class sizes follow $N_c = N_{\max}\rho^{c/(K-1)}$, $c \in \{0, \dots, K-1\}$, where $K = 10$, $N_{\max}$ is the largest class size, and $\rho = N_{\min}/N_{\max}$ is the ratio of the smallest to largest class sizes. We consider mild imbalance with $\rho = 0.1$ and severe imbalance with $\rho = 0.01$. The original class-balanced test set is retained, and SVHN and CIFAR-100 are used as OOD datasets. This setting evaluates the predictor and its uncertainty estimates when training information is unevenly distributed across classes. After constructing the long-tailed training set, we apply the same stratified 95:5 split used for the other CIFAR benchmarks to obtain the training and development subsets.

SVHN. This dataset (Netzer et al., 2011) contains cropped images of house-number digits from Google Street View. It comprises 73,257 training images, 26,032 test images, and an additional set of 531,131 images. Each image is an RGB image of size 32×32. Its digit-recognition content differs from the animal and object categories of the CIFAR benchmarks while retaining the same image dimensions. We use only the official test set for OOD evaluation with CIFAR-10, CIFAR-100, and CIFAR-10-LT; neither its training split nor its additional images are used to fit the evaluated ID models.

Tiny ImageNet-200. This dataset (Le & Yang, 2015) is a reduced-resolution ImageNet benchmark containing 200 object categories. It provides 100,000 training images, 10,000 validation images, and 10,000 test images at a resolution of 64×64. We use the official `test` partition in full as an OOD benchmark for the CIFAR-100 classifier. All 10,000 test images are resized to 32×32 to match the classifier's input resolution. Because OOD detection requires only distinguishing ID from OOD inputs, the test labels are not used. The `train` and `val` partitions are not used for model selection, post-hoc fitting, or evaluation.

D.1.2 PRETRAINED SETTING

ImageNet-1K. This dataset (Deng et al., 2009; Russakovsky et al., 2015) is a large-scale natural-image benchmark with 1,000 object categories and 50,000 labeled validation images. We use it as the ID dataset for pretrained vision experiments, starting from a ResNet-50 initialized with the TorchVision ImageNet-1K V2 weights. The official validation set is divided into an adaptation subset and an ID evaluation subset using a 20:80 split. The former supplies data for adaptation and calibration, while the latter is reserved for evaluation. The underlying predictor remains frozen during BEDI fitting.

ImageNet-A. This dataset (Hendrycks et al., 2021b) contains naturally occurring, unmodified images that are challenging for ImageNet classifiers. Unlike synthetic corruptions, these examples arise from real-world image variation rather than an explicitly applied perturbation. We use it to evaluate whether uncertainty distinguishes shifted inputs from held-out ImageNet ID images. No ImageNet-A example is used for adaptation, calibration, or uncertainty fitting.

ImageNet-Sketch. This dataset (Wang et al., 2019) contains approximately 50,000 sketch images corresponding to the 1,000 ImageNet categories. The shift from natural photographs to sketches alters appearance and texture while retaining the underlying object categories. We use it as a distribution-shift benchmark for the pretrained ImageNet classifier without fitting on sketch data.

ImageNet-R. This dataset (Hendrycks et al., 2021a) contains 30,000 renditions of 200 ImageNet categories, including paintings, cartoons, embroidery, sculptures, and toys. These examples provide a complementary shift from photographic images to alternative depictions of familiar object categories. We use it only for distribution-shift evaluation. Results are reported separately for ImageNet-A, ImageNet-Sketch, and ImageNet-R, together with their average.

Dry Bean. This dataset (Koklu & Ozkan, 2020) contains 13,611 instances from seven dry-bean varieties. Each instance is represented by 16 numerical measurements extracted from a bean image, including size and shape descriptors. We use the pretrained TabPFN-3 classifier as the underlying predictor. Three class-held-out OOD settings are constructed by excluding BARBUNYA, DERMASON, and SIRA in turn. In each setting, the excluded variety is treated as OOD, while the remaining six varieties constitute the ID classification problem. The held-out class is excluded from all ID training, calibration, and uncertainty-fitting data.

OpenBookQA. This dataset (Mihaylov et al., 2018) is a multiple-choice question-answering benchmark built around elementary science knowledge. Its questions require combining scientific facts with broader common knowledge rather than merely locating an answer in a supplied passage. Each question has four candidate answers. We use it as an ID language benchmark, adapting Llama 3.1 8B to the ID training split with a LoRA adapter before freezing the predictor for BEDI. OOD evaluation uses the selected MMLU subjects described below.

MMLU. This benchmark (Hendrycks et al., 2020) evaluates multiple-choice question answering across 57 academic and professional subjects. For language OOD evaluation, we use eight subjects from `cais/mmlu`, taking each subject's full `test` partition: `college_computer_science`, `computer_security`, `high_school_computer_science`, and `machine_learning` from computer science; `electrical_engineering` from engineering; and `college_mathematics`, `high_school_mathematics`, and `abstract_algebra` from mathematics. The resulting OOD set contains 1,027 questions in total. These examples are used exclusively for evaluation and are not used for LoRA checkpoint selection, calibration, or dispersion fitting. No examples from the `dev`, `validation`, or `auxiliary_train` partitions are used. Here, OOD refers to the downstream ID task rather than implying that the language model did not encounter related content during pretraining.

D.2 Evaluation Protocols

Data partitioning. For the trained-from-scratch CIFAR benchmarks, the ID training data are split into training and development subsets using a fixed stratified 95:5 split. The training subset is used to train the classifier and construct the support and Laplace quantities, while the development subset is used for checkpoint selection, temperature calibration, and post-hoc fitting of BEDI. The official test sets are used only for evaluation. The pretrained settings follow their dataset-specific protocols. ImageNet uses the 20:80 adaptation–evaluation split described in Section D.1.2; Dry Bean holds out the designated OOD class before forming the ID partitions; and the language experiments use only ID data for adaptation and development-set fitting. No OOD or shifted examples are used for model selection, calibration, or dispersion inference.

Misclassification detection. Correctly classified ID examples are treated as positives and misclassified examples as negatives. For BEDI, the confidence score is the negative total predictive uncertainty,

$$s_{\mathrm{mis}}(\mathbf{x}) = -\,\mathrm{TU}(\mathbf{x}) = -H(\widetilde{\mathbf{m}}(\mathbf{x}))\,.$$

Because BEDI preserves the calibrated predictive mean, dispersion inference does not alter this score relative to the same calibrated predictor. For each baseline, we use the misclassification-detection score recommended in its original paper when specified; otherwise, we use negative predictive entropy.

OOD and shift detection. ID examples are treated as positives and OOD or shifted examples as negatives. For BEDI, we use negative epistemic uncertainty as the ID-confidence score,

$$s_{\mathrm{ID}}(\mathbf{x}) = -\,\mathrm{EU}(\mathbf{x}),$$

where EU is defined by the entropy decomposition in Equation (11). AUPR is the primary detection metric, with values reported on a 0–100 scale. For each baseline, we use its recommended OOD-detection score, orienting it so that higher values indicate ID examples.

For methods evaluated under our shared protocol, we use the same data partitions and preprocessing. The results for DMM and ETN on CIFAR-10 are taken from Chun et al. (2026), while the results for the EDL-based methods on CIFAR-10, CIFAR-100, and CIFAR-10-LT are taken from their corresponding papers (Deng et al., 2023; Yoon & Kim, 2024; Chen et al., 2024; 2025; Yoon & Kim, 2026b;a; Ni et al., 2026). All other results are reproduced under our protocol.

We report results over five random seeds for each from-scratch setting and ImageNet, fifteen seeds for each of the three Dry Bean held-out classes, and three seeds for OpenBookQA.

D.3 Implementation Details

Table 2: **(Implementation settings)** Predictor and BEDI settings. For the from-scratch predictors, patience is counted in evaluation events, with evaluation every five epochs. The RBF bandwidth multiplier is selected using ID data; the KL weight β is fixed as specified below. The values of M, N_{LA}, and J are computational approximation budgets.

Predictor settings

Dataset	Predictor	Initial LR	Early-stopping patience
CIFAR-10	VGG-16	0.05	13
CIFAR-100	ResNet-18	0.12	17
CIFAR-10-LT	VGG-16	0.05	15
ImageNet	ResNet-50 (ImageNet-1K V2)		Frozen pretrained predictor
Dry Bean	TabPFN-3		Frozen pretrained predictor
OpenBookQA	Llama 3.1 8B + LoRA	10^{-4}	3 epochs

From-scratch optimizer	SGD, momentum 0.9
From-scratch LR schedule	5-epoch warm-up; decay $\times 0.1$ every 82 epochs
Maximum training epochs	250
Spectral normalization	Enabled
OpenBookQA optimizer	AdamW

BEDI dataset-specific settings

ID dataset	Predictor	Bandwidth multiplier	M	N_{LA}
CIFAR-10	VGG-16	5	200	1,000
CIFAR-100	ResNet-18	3	200	1,000
CIFAR-10-LT	VGG-16	5	200	1,000
ImageNet	ResNet-50	15	2,000	1,000
Dry Bean	TabPFN-3	1	128	All
OpenBookQA	Llama 3.1 8B + LoRA	80	1,000	1,000

Other BEDI settings

KL weight β	0.3 (Dry Bean: 1)
Observation-noise variance σ_n^2	1
Laplace class-weight prior	$\mathcal{N}(\mathbf{0}, 0.1 I_D)$
Laplace covariance structure	Full within-class (ImageNet: diagonal)
Fisher-curvature scale σ_{curv}^2	1
Dispersion bounds $d_{\min}, d_{\max}$	10^{-4}, 0.95
Global prior $p(\boldsymbol{\theta})$	$\mathcal{N}(\boldsymbol{\theta}_0, 0.1^2 I)$, $\boldsymbol{\theta}_0 = (-2, \mathbf{0}_D, \mathrm{softplus}^{-1}(1))$
$\tau_0, \sigma_{\mathrm{floor}}$	0.71, 0.71
Variational posterior	Mean-field Gaussian
Optimizer / steps / learning rate	Adam / 200 / 10^{-2}
Posterior predictive draws J	64

Predictors. For the from-scratch experiments, we use VGG-16 for CIFAR-10 and CIFAR-10-LT and ResNet-18 for CIFAR-100. The classifiers are trained with cross-entropy using SGD with momentum 0.9 and spectral normalization. The initial learning rates are 0.05, 0.12, and 0.05 for CIFAR-10, CIFAR-100, and CIFAR-10-LT, respectively. We use a 5-epoch linear warm-up followed by stepwise learning-rate decay by a factor of 0.1 every 82 epochs, with a maximum of 250 epochs. Models are

evaluated every 5 epochs, and early stopping uses development top-1 accuracy with patience 13, 17, and 15, respectively; development NLL is used to break ties.

For the pretrained experiments, we use the TorchVision ResNet-50 with ImageNet-1K V2 weights and its original classification head without additional predictor training, and TabPFN-3 as a frozen predictor for Dry Bean. For OpenBookQA, Llama 3.1 8B is adapted to the ID task with a LoRA adapter trained for three epochs using cross-entropy and AdamW with learning rate 10^{-4}. In all settings, the predictor is frozen before fitting BEDI.

Calibration and support estimation. We apply scalar temperature scaling (Guo et al., 2017) to the frozen predictor by minimizing development-set NLL with L-BFGS. The resulting calibrated predictive mean is fixed throughout dispersion inference. Support is estimated from ℓ_2-normalized frozen representations using the Nyström approximation described in Section 3.2.1. Inducing centroids are obtained by class-wise k-means, and the RBF bandwidth is defined as the median nearest-centroid distance multiplied by the dataset-specific factor in Table 2. The auxiliary Bayesian linear model uses a standard Gaussian coefficient prior, and the resulting effective sample size is normalized by its training-set average.

Laplace construction. The class-factorized last-layer Laplace approximation uses the scale-normalized curvature derived in Section A.2 and is centered at the temperature-scaled classifier weights, with the bias fixed. The curvature weights and softmax Jacobian are evaluated at the original calibrated probabilities, while normalized representations are used to construct the curvature and evaluate disagreement. We fix the Laplace class-weight prior to $\mathcal{N}(\mathbf{0}, 0.1 I_D)$ and the Fisher-curvature scale to $\sigma^2_{\text{curv}} = 1$ in all experiments. We retain full $D \times D$ within-class covariance blocks except on ImageNet, where we use diagonal blocks. Cross-class covariance blocks are discarded in all settings. We project the resulting disagreement covariance onto the fixed-mean Dirichlet family to construct development-set pseudo-observations. The projected dispersion is clipped to $[d_{\min}, d_{\max}]$, where $d_{\min} = 10^{-4}$ and $d_{\max} = 0.95$, before the logit transformation. The Laplace covariances and pseudo-observations are used only during post-hoc fitting and are discarded before test-time prediction.

Dispersion inference. For the global prior $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}_0, \sigma_0^2 I)$ defined in Section 3.2.1, we use

$$\boldsymbol{\theta}_0 = \left(-2, \mathbf{0}_D, \text{softplus}^{-1}(1)\right), \qquad \sigma_0 = 0.1.$$

We approximate the posterior over the shared dispersion parameters with a mean-field Gaussian and optimize the collapsed variational objective with Adam. At evaluation, posterior predictive draws are used to compute the entropy-based uncertainty decomposition. Expected categorical entropy conditional on each dispersion draw is evaluated analytically, without sampling class-probability vectors, and the neural representation is computed only once per input.

Hyperparameters and computational budgets. We distinguish selected hyperparameters, fixed model settings, and computational approximation budgets.

Selected hyperparameters. We select the dataset-specific RBF bandwidth multiplier using only ID data, checking the relative Frobenius error of the Nyström kernel approximation and the fraction of queries whose estimated posterior variance exceeds the prior variance. The selected multiplier is fixed across random seeds.

Fixed model settings. We set the KL weight to $\beta = 0.3$, except for Dry Bean, whose separate variational implementation uses $\beta = 1$. We do not tune β for downstream performance. For the from-scratch experiments, the KL weight is linearly increased from zero to its fixed value over the first half of the 200 dispersion-head optimization steps. We set the auxiliary observation-noise variance to $\sigma_n^2 = 1$. We also set $\tau_0 = \sigma_{\text{floor}} = 0.71$, approximately $1/\sqrt{2}$, assigning equal base variance to the prior and pseudo-likelihood components on an approximately unit scale.

Computational budgets. The numbers of inducing centroids M, representations used for Laplace construction N_{LA}, and posterior predictive draws J control approximation cost and are not tuned for downstream performance. The settings are summarized in Table 2. OOD and shifted evaluation data are never used for model selection, calibration, or hyperparameter selection.

D.4 Computational Cost and Scalability

Fitting cost. BEDI is fitted post hoc to a frozen classifier, without retraining the predictive backbone. Given N_{tr} frozen representations of dimension D and M inducing centroids, the Nyström kernel computations and construction of the auxiliary posterior covariance require $O(N_{\mathrm{tr}}MD + N_{\mathrm{tr}}M^2 + M^3)$ time. For N_{LA} representations and K classes, constructing the full within-class Laplace covariance blocks requires $O(N_{\mathrm{LA}}KD^2 + KD^3)$ time and $O(KD^2)$ temporary covariance storage. The Laplace covariances are discarded after constructing the disagreement-derived pseudo-observations.

On CIFAR-10 with VGG-16, the complete post-hoc procedure, including temperature scaling, support and Laplace construction, and variational fitting of dispersion, takes 41 s. This is longer than LLLA (2 s), but substantially shorter than DMM (1,128 s) and ETN (931 s). The end-to-end entries marked by $*$ in Table 3 estimate optimization costs from median step times and training schedules and serve as computational references rather than directly comparable post-hoc fitting costs.

Table 3: **(Computational cost)** Costs on CIFAR-10 with VGG-16, measured on one RTX 3090. Post-hoc fitting times exclude base-classifier training. Entries marked by $*$ estimate end-to-end optimization costs from median step times and training schedules. Inference latency uses batch size one, while peak inference memory uses batch size 128. Extra parameters and buffers denote persistent state beyond the base classifier; counts are numbers of scalar entries.

		Fitting	Persistent state		Inference	
Method	Fit type	Fit (s)	Extra parameters	Extra buffers	Latency (ms)	Peak memory (GB)
MSP	None	–	0	0	3.12 ± 0.03	0.218
EDL*	E2E	3,656	0	0	3.15 ± 0.03	0.218
$\mathcal{I}$-EDL*	E2E	3,656	0	0	3.11 ± 0.03	0.218
R-EDL*	E2E	3,656	0	0	3.20 ± 0.03	0.218
Re-EDL*	E2E	3,550	0	0	3.15 ± 0.03	0.218
$\mathcal{F}$-EDL*	E2E	3,723	267K	0	3.79 ± 0.01	0.219
MoDEX*	E2E	3,767	269K	0	4.06 ± 0.10	0.219
DMM	Post-hoc	1,128	186K	0	3.70 ± 0.00	0.218
ETN	Post-hoc	931	66K	0	3.63 ± 0.03	0.218
SNGP	E2E	3,142	5K	2,622K	3.67 ± 0.03	0.226
LLLA	Post-hoc	2	0	5K	4.43 ± 0.01	0.218
VBLL*	E2E	4,358	5K	0	4.14 ± 0.04	0.217
MC Dropout*	E2E	3,517	0	0	30.74 ± 0.52	0.219
BEDI	Post-hoc	41	1K	183K	5.03 ± 0.06	0.218

Inference cost. At test time, BEDI requires one forward pass through the frozen classifier, followed by kernel computations and posterior predictive sampling of dispersion. Given the resulting representation, the Nyström and auxiliary-variance computations cost $O(MD + M^2)$. For J posterior predictive draws, evaluating the sampled prior means costs $O(JD)$ and computing the conditional entropy terms costs $O(JK)$. These operations require no further backbone evaluations.

With $M = 200$ and $J = 64$ on CIFAR-10, inference takes 5.03 ± 0.06 ms per input, compared with 3.12 ± 0.03 ms for MSP and 3.15 ± 0.03 ms for EDL. The latency is moderately higher than LLLA (4.43 ± 0.01 ms) and VBLL (4.14 ± 0.04 ms), but substantially lower than ten-pass MC Dropout (30.74 ± 0.52 ms). Peak inference memory is 0.218 GB at batch size 128. Beyond the base classifier, BEDI retains approximately 1K additional parameters and 183K buffer entries. Thus, BEDI incurs additional computation to estimate epistemic uncertainty while evaluating the backbone only once per input.

E Additional Experimental Results

E.1 From-Scratch Setting

Tables 4 and 5 extend the main-text comparisons to additional baselines, CIFAR-100, and both class-imbalance ratios. BEDI attains the highest reported mean test accuracy in all four settings, although the margins over the closest baselines are small. Because dispersion is inferred after classifier training, it retains the calibrated base classifier's predictions. It also leads in misclassification AUPR on CIFAR-10 and CIFAR-10-LT at $\rho = 0.1$, while MoDEX leads on CIFAR-100 and CIFAR-10-LT at $\rho = 0.01$.

For OOD detection, BEDI achieves the highest AUPR in seven of the eight evaluated ID–OOD pairs. It leads on both OOD datasets for CIFAR-10 and at each CIFAR-10-LT imbalance ratio; at $\rho = 0.01$, its AUPRs reach 81.25 on SVHN and 77.39 on CIFAR-100. On CIFAR-100, it leads on SVHN (94.68), whereas MoDEX performs better on Tiny ImageNet (81.76 versus 75.87). These results are consistent with support-sensitive dispersion inference, which combines effective sample size with Bayesian predictive disagreement while leaving the predictive mean unchanged. The temperature-scaled softmax classifier used here is one possible base predictor; in principle, BEDI could instead be fitted to another classifier selected for stronger accuracy or misclassification detection.

Table 4: **(Standard Setting)** UQ-related downstream tasks on CIFAR-10 and CIFAR-100. OOD AUPR is reported for the datasets in each header; C-100 and TIN denote CIFAR-100 and Tiny ImageNet, respectively.

Method	CIFAR-10			CIFAR-100		
	Test Acc.	Miscl. AUPR	OOD AUPR SVHN | C-100	Test Acc.	Miscl. AUPR	OOD AUPR SVHN | TIN
Dropout	90.16 ± 0.2	98.86 ± 0.6	78.40 ± 3.9 | 85.39 ± 0.6	65.94 ± 0.6	92.00 ± 0.3	71.83 ± 2.0 | 74.93 ± 0.6
LLLA	91.97 ± 0.2	98.17 ± 0.5	70.79 ± 9.8 | 76.03 ± 2.7	72.26 ± 0.3	94.68 ± 0.2	63.10 ± 3.5 | 78.96 ± 0.3
VBLL	92.50 ± 0.8	98.52 ± 0.2	81.30 ± 13.9 | 80.55 ± 4.2	73.70 ± 0.3	94.69 ± 0.2	69.40 ± 1.0 | 78.45 ± 0.5
SNGP	92.23 ± 0.7	98.90 ± 0.1	87.23 ± 2.1 | 81.53 ± 2.1	72.49 ± 0.2	94.38 ± 0.1	74.68 ± 4.6 | 78.11 ± 0.3
EDL	88.48 ± 0.3	98.74 ± 0.7	82.32 ± 1.2 | 87.13 ± 0.3	45.91 ± 5.6	91.28 ± 0.8	56.21 ± 3.1 | 70.13 ± 2.0
$\mathcal{I}$-EDL	89.20 ± 0.3	98.72 ± 0.1	82.96 ± 2.2 | 84.84 ± 0.6	66.38 ± 0.5	92.84 ± 0.1	67.51 ± 2.9 | 75.86 ± 0.3
DAEDL	91.11 ± 0.2	99.08 ± 0.0	85.54 ± 1.4 | 88.19 ± 0.1	66.01 ± 2.6	86.00 ± 0.3	72.07 ± 4.1 | 77.40 ± 1.6
R-EDL	90.09 ± 0.3	98.98 ± 0.1	85.00 ± 1.2 | 87.73 ± 0.3	63.53 ± 0.5	92.69 ± 0.2	61.80 ± 3.4 | 69.78 ± 1.3
Re-EDL	90.09 ± 0.3	98.81 ± 0.1	89.94 ± 1.4 | 88.31 ± 0.2	64.53 ± 0.7	92.63 ± 0.4	68.37 ± 2.8 | 76.87 ± 0.7
$\mathcal{F}$-EDL	91.19 ± 0.2	99.10 ± 0.0	91.20 ± 1.3 | 88.37 ± 0.3	69.40 ± 0.2	94.01 ± 0.1	75.35 ± 2.3 | 80.58 ± 0.2
DAPPr	92.00 ± 0.2	99.23 ± 0.0	91.72 ± 1.2 | 89.39 ± 0.3	70.85 ± 0.2	94.39 ± 0.1	73.32 ± 3.3 | 79.11 ± 0.1
MoDEX	92.46 ± 0.2	99.18 ± 0.0	91.58 ± 0.4 | 89.28 ± 0.3	75.91 ± 0.9	**96.17** ± 0.2	77.90 ± 1.5 | **81.76** ± 0.3
DMM	87.10 ± 3.2	98.46 ± 0.2	81.59 ± 7.0 | 84.61 ± 2.5	75.36 ± 0.3	94.73 ± 0.1	69.90 ± 6.4 | 78.89 ± 0.4
ETN	90.70 ± 0.0	98.99 ± 0.1	85.60 ± 1.5 | 86.84 ± 0.3	76.43 ± 0.3	93.25 ± 0.2	44.42 ± 13.02 | 62.28 ± 0.6
BEDI	**92.58** ± 0.5	**99.36** ± 0.1	**93.10** ± 1.2 | **90.35** ± 0.6	**76.60** ± 0.4	95.88 ± 0.1	**94.68** ± 1.1 | 75.87 ± 0.4

Table 5: **(Long-tailed Setting)** UQ-related downstream tasks on CIFAR-10-LT under mild ($\rho = 0.1$) and severe ($\rho = 0.01$) class imbalance.

Method	CIFAR-10-LT ($\rho = 0.1$)			CIFAR-10-LT ($\rho = 0.01$)		
	Test Acc.	Miscl. AUPR	OOD AUPR SVHN | C-100	Test Acc.	Miscl. AUPR	OOD AUPR SVHN | C-100
Dropout	70.87 ± 3.0	89.82 ± 2.3	37.37 ± 1.4 | 61.18 ± 1.3	39.22 ± 3.1	63.62 ± 2.7	33.33 ± 1.7 | 54.17 ± 1.1
LLLA	84.61 ± 0.5	94.17 ± 0.3	56.76 ± 4.3 | 69.72 ± 0.8	66.46 ± 1.1	76.71 ± 2.3	44.83 ± 3.8 | 60.34 ± 2.2
VBLL	85.12 ± 0.5	89.96 ± 0.5	42.93 ± 7.7 | 59.24 ± 1.0	66.30 ± 1.8	71.90 ± 1.8	37.62 ± 13.8 | 55.36 ± 3.1
SNGP	84.69 ± 0.6	93.81 ± 0.2	68.49 ± 7.6 | 71.70 ± 1.2	66.77 ± 1.1	75.60 ± 1.3	50.64 ± 4.2 | 62.62 ± 0.7
EDL	79.09 ± 0.4	95.36 ± 0.1	72.18 ± 2.1 | 80.09 ± 0.7	42.62 ± 2.7	82.63 ± 1.7	51.99 ± 3.8 | 66.86 ± 0.9
$\mathcal{I}$-EDL	84.86 ± 0.1	97.31 ± 0.2	79.83 ± 3.9 | 83.50 ± 0.4	57.88 ± 1.3	84.10 ± 1.3	52.85 ± 6.8 | 69.19 ± 1.3
DAEDL	84.95 ± 0.4	95.22 ± 0.4	69.40 ± 4.5 | 74.56 ± 1.7	63.36 ± 1.4	82.15 ± 1.0	51.03 ± 5.6 | 65.31 ± 1.2
R-EDL	85.35 ± 0.2	94.35 ± 0.2	60.58 ± 5.0 | 69.53 ± 1.6	63.36 ± 1.0	78.34 ± 1.0	48.71 ± 7.1 | 64.20 ± 1.4
Re-EDL	84.43 ± 0.8	94.75 ± 0.7	59.08 ± 10.9 | 72.37 ± 3.5	66.27 ± 0.4	83.40 ± 1.2	39.48 ± 9.4 | 62.51 ± 2.4
$\mathcal{F}$-EDL	85.46 ± 0.2	97.60 ± 0.1	85.36 ± 1.5 | 83.64 ± 0.7	63.73 ± 1.4	85.99 ± 1.7	62.56 ± 2.8 | 70.18 ± 2.0
DAPPr	86.37 ± 0.3	97.69 ± 0.1	85.63 ± 2.0 | 84.32 ± 0.5	68.81 ± 0.9	87.01 ± 1.3	64.80 ± 3.6 | 71.77 ± 1.3
MoDEX	86.28 ± 0.5	97.78 ± 0.1	86.60 ± 0.7 | 84.22 ± 0.2	71.53 ± 1.1	**91.90** ± 1.2	72.05 ± 3.5 | 76.52 ± 0.9
DMM	87.33 ± 0.5	97.84 ± 0.2	81.70 ± 5.3 | 83.56 ± 0.7	71.52 ± 0.9	90.15 ± 0.6	74.38 ± 2.6 | 75.81 ± 0.6
ETN	87.35 ± 0.6	97.49 ± 0.4	70.05 ± 6.5 | 76.35 ± 1.8	71.21 ± 1.3	82.06 ± 1.5	49.50 ± 10.0 | 58.13 ± 2.2
BEDI	**87.37** ± 0.7	**98.10** ± 0.2	**88.16** ± 2.4 | **85.39** ± 0.7	**72.04** ± 0.9	91.29 ± 0.6	**81.25** ± 1.8 | **77.39** ± 1.0

E.2 Pretrained Setting

Tables 6 to 8 report the full results for the pretrained tabular, vision, and language settings. Across these settings, BEDI preserves the base predictor's calibrated probabilities while achieving strong OOD detection. On Dry Bean, it matches temperature-scaled TabPFN in accuracy and misclassification AUPR and increases the average class-held-out OOD AUPR from 89.55 to 91.71. On ImageNet, it retains the highest reported accuracy and achieves the best OOD AUPR on all three shifts, with an average of 88.62 versus 87.56 for the strongest baseline. On OpenBookQA, its OOD AUPR reaches 91.78 versus 78.85 for the strongest baseline, while its misclassification AUPR remains within 0.2 points of the best result. These results show that post-hoc dispersion inference can improve OOD detection without changing the calibrated predictive probabilities or classification decisions.

The language setting makes this distinction particularly clear. MoDEX and R-EDL improve OpenBookQA accuracy over the cross-entropy LoRA backbone by 1.3 and 1.1 points, respectively, but their OOD AUPRs are 37.67 and 75.37, below the backbone's 78.85 and BEDI's 91.78. MoDEX, the most accurate adapter, performs especially poorly at detecting MMLU inputs. These results suggest that jointly improving ID classification and fitting evidence does not necessarily produce an informative epistemic score under shift. BEDI instead preserves the calibrated predictive mean and fits dispersion separately, retaining the base predictor's accuracy while substantially improving OOD detection.

Table 6: **(Pretrained Tabular Setting)** UQ-related downstream tasks on Dry Bean with a pretrained TabPFN-3. OOD AUPR is reported for the classes in each header; each column holds that class out from training and treats it as OOD, and Avg. is their mean. $\mathcal{I}$-EDL, R-EDL, and MoDEX are fine-tuned from the pretrained predictor, whereas the remaining methods are applied post hoc. Accuracy differs across rows because each method attaches to a different predictive mean: the post-hoc baselines read a calibrated linear probe over TabPFN's frozen embeddings, whereas BEDI centers on TabPFN's own probability vector.

	Classification		OOD AUPR			
Method	Test Acc.	Miscl. AUPR	BARBUNYA	DERMASON	SIRA	Avg.
TabPFN	**95.61** $\pm$ 0.2	**99.66** $\pm$ 0.0	90.11 $\pm$ 1.6	85.31 $\pm$ 0.6	93.09 $\pm$ 0.2	89.50 $\pm$ 0.6
TabPFN + TS	**95.61** $\pm$ 0.2	**99.66** $\pm$ 0.0	90.15 $\pm$ 1.6	85.34 $\pm$ 0.6	93.18 $\pm$ 0.2	89.55 $\pm$ 0.6
LLLA	95.49 $\pm$ 0.2	99.58 $\pm$ 0.0	94.76 $\pm$ 1.2	85.88 $\pm$ 1.2	90.04 $\pm$ 0.6	90.23 $\pm$ 0.3
VBLL	95.48 $\pm$ 0.2	99.58 $\pm$ 0.0	**95.28** $\pm$ 0.4	80.86 $\pm$ 3.4	86.63 $\pm$ 0.9	87.59 $\pm$ 1.3
$\mathcal{I}$-EDL	95.45 $\pm$ 0.1	**99.66** $\pm$ 0.0	51.05 $\pm$ 1.1	22.86 $\pm$ 0.8	27.78 $\pm$ 0.0	33.90 $\pm$ 0.3
R-EDL	95.32 $\pm$ 0.1	99.20 $\pm$ 0.1	57.54 $\pm$ 0.9	31.87 $\pm$ 0.6	40.34 $\pm$ 1.1	43.25 $\pm$ 0.5
MoDEX	95.06 $\pm$ 0.2	98.66 $\pm$ 0.2	84.52 $\pm$ 7.3	67.37 $\pm$ 15.8	59.09 $\pm$ 8.3	70.32 $\pm$ 3.5
DMM	95.41 $\pm$ 0.1	99.58 $\pm$ 0.0	90.14 $\pm$ 7.8	81.00 $\pm$ 5.2	92.76 $\pm$ 2.0	87.97 $\pm$ 3.8
ETN	95.49 $\pm$ 0.2	99.46 $\pm$ 0.1	90.26 $\pm$ 1.8	84.36 $\pm$ 3.1	85.26 $\pm$ 4.3	86.63 $\pm$ 1.4
BEDI	**95.61** $\pm$ 0.2	**99.66** $\pm$ 0.0	94.35 $\pm$ 1.1	**86.19** $\pm$ 0.8	**94.59** $\pm$ 0.3	**91.71** $\pm$ 0.4

E.3 Detection AUROC

The main results use AUPR, which depends on the prevalence of positive examples. We additionally report AUROC, whose chance level is 0.5 regardless of prevalence (Tables 9 and 10). The main OOD and shift detection result is consistent under AUROC: BEDI achieves the highest score in all seven settings. It also achieves the highest misclassification AUROC in five settings; LLLA leads by 0.11 points on ImageNet and 0.23 points on OpenBookQA.

E.4 In-Distribution Calibration

We evaluate in-distribution predictive probabilities using expected calibration error (ECE) and negative log-likelihood (NLL), which assess confidence calibration and probabilistic prediction quality, respectively. We compare BEDI with LLLA, VBLL, MoDEX, DMM, and ETN across from-scratch and pretrained settings.

As shown in Tables 11 and 12, BEDI achieves the lowest ECE in five of seven settings and the lowest NLL in six. It obtains lower ECE and NLL than MoDEX in every setting. Together with

Table 7: **(Pretrained Vision Setting)** UQ-related downstream tasks on ImageNet with a pretrained ResNet-50. Shift AUPR is reported for the datasets in each header; IN-A, IN-Sketch, and IN-R denote ImageNet-A, ImageNet-Sketch, and ImageNet-R, respectively, and Avg. is their mean. $\mathcal{I}$-EDL, R-EDL, and MoDEX train an evidential head on the same frozen features, whereas the remaining methods are applied post hoc.

	Classification		Shift AUPR			
Method	Test Acc.	Miscl. AUPR	IN-A	IN-Sketch	IN-R	Avg.
LLLA	80.25 ± 0.0	**95.72** ± 0.1	96.91 ± 0.1	75.30 ± 0.3	83.55 ± 0.2	85.25 ± 0.2
VBLL	80.25 ± 0.0	95.70 ± 0.1	95.92 ± 0.1	70.67 ± 0.4	80.45 ± 0.2	82.35 ± 0.2
$\mathcal{I}$-EDL	80.23 ± 0.1	89.85 ± 0.0	96.98 ± 0.0	78.31 ± 0.1	87.41 ± 0.0	87.56 ± 0.0
R-EDL	80.21 ± 0.0	89.82 ± 0.0	96.57 ± 0.0	77.51 ± 0.0	86.75 ± 0.0	86.94 ± 0.0
MoDEX	80.13 ± 0.1	89.44 ± 0.1	96.36 ± 0.0	78.24 ± 0.0	87.08 ± 0.0	87.23 ± 0.0
DMM	70.92 ± 0.1	87.95 ± 0.3	92.66 ± 0.2	66.07 ± 1.5	76.78 ± 1.4	78.50 ± 0.9
ETN	80.24 ± 0.0	92.11 ± 0.0	96.14 ± 0.1	72.64 ± 1.1	81.37 ± 0.8	83.38 ± 0.6
BEDI	**80.26** ± 0.0	95.66 ± 0.0	**97.24** ± 0.1	**81.03** ± 0.4	**87.60** ± 0.2	**88.62** ± 0.2

Table 8: **(Pretrained Language Setting)** UQ-related downstream tasks on OpenBookQA with Llama 3.1 8B, using selected MMLU subjects as OOD data. OOD AUPR is averaged over the evaluated subjects. Post-hoc methods use a frozen cross-entropy-trained LoRA backbone; $\mathcal{I}$-EDL, R-EDL, and MoDEX train their adapters using their respective objectives. BEDI preserves the calibrated class probabilities of the frozen predictor.

Method	Test Acc.	Miscl. AUPR	OOD AUPR
LoRA (MSP) + TS	86.53 ± 0.3	97.38 ± 0.6	78.85 ± 1.1
MC-Dropout	86.47 ± 0.4	97.61 ± 0.6	71.67 ± 2.2
LLLA	86.67 ± 0.7	**97.76** ± 0.6	67.70 ± 3.9
VBLL	86.60 ± 0.5	96.98 ± 0.6	74.74 ± 4.3
$\mathcal{I}$-EDL	87.13 ± 0.8	96.52 ± 0.6	74.25 ± 7.1
R-EDL	87.67 ± 0.3	94.41 ± 1.0	75.37 ± 3.0
MoDEX	**87.80** ± 0.4	97.48 ± 0.9	37.67 ± 0.4
DMM	86.60 ± 0.2	96.33 ± 0.4	75.91 ± 2.9
ETN	86.53 ± 0.3	94.82 ± 0.7	44.43 ± 7.8
BEDI	86.53 ± 0.3	97.61 ± 0.7	**91.78** ± 7.2

Table 9: **(Detection AUROC: from-scratch setting)** Misclassification and OOD detection AUROC over five seeds. OOD AUROC averages the two OOD datasets for each setting.

	CIFAR-10		CIFAR-100		CIFAR-10-LT ($\rho = 0.1$)		CIFAR-10-LT ($\rho = 0.01$)	
Method	Miscl.	OOD	Miscl.	OOD	Miscl.	OOD	Miscl.	OOD
LLLA	87.65 ± 0.5	83.33 ± 1.1	86.99 ± 0.4	76.46 ± 0.9	79.41 ± 0.8	74.12 ± 1.3	66.21 ± 2.0	62.85 ± 1.4
VBLL	88.40 ± 1.0	86.30 ± 0.8	86.87 ± 0.2	78.01 ± 0.4	64.26 ± 1.9	61.08 ± 3.1	56.71 ± 4.9	54.82 ± 6.3
SNGP	90.43 ± 0.5	87.26 ± 0.5	86.24 ± 0.3	81.14 ± 1.4	77.19 ± 0.7	77.72 ± 1.4	61.95 ± 1.1	62.06 ± 1.3
MoDEX	91.29 ± 0.7	88.43 ± 1.2	85.09 ± 0.2	76.34 ± 3.0	85.56 ± 1.1	82.46 ± 3.0	74.39 ± 2.0	72.40 ± 5.7
DMM	92.44 ± 0.5	90.14 ± 0.9	85.97 ± 0.1	76.23 ± 1.4	87.41 ± 0.6	86.25 ± 2.2	76.98 ± 1.2	77.90 ± 2.5
ETN	91.85 ± 0.6	84.55 ± 2.1	82.92 ± 0.5	63.30 ± 1.8	86.63 ± 1.1	76.86 ± 6.8	65.85 ± 1.1	62.85 ± 6.0
BEDI	**93.00** ± 0.5	**92.43** ± 0.4	**87.57** ± 0.3	**84.37** ± 1.0	**88.40** ± 0.4	**87.72** ± 1.2	**79.94** ± 0.8	**79.81** ± 0.5

Table 10: **(Detection AUROC: pretrained setting)** Misclassification and shift/OOD detection AUROC. Shift/OOD AUROC averages three ImageNet shifts or three Dry Bean held-out classes; OpenBookQA uses the evaluated MMLU subjects. Results use five ImageNet seeds, fifteen Dry Bean seeds per held-out class, and three OpenBookQA seeds. Scores follow the conventions in Table 9.

	ImageNet		Dry Bean		OpenBookQA	
Method	Miscl.	Shift/OOD	Miscl.	Shift/OOD	Miscl.	Shift/OOD
LLLA	**86.55** $\pm$ 0.1	83.33 $\pm$ 0.2	93.10 $\pm$ 1.2	86.23 $\pm$ 6.0	**87.77** $\pm$ 2.2	72.31 $\pm$ 3.3
VBLL	84.22 $\pm$ 0.2	79.91 $\pm$ 0.3	93.09 $\pm$ 1.3	89.41 $\pm$ 3.0	84.79 $\pm$ 2.5	85.05 $\pm$ 5.0
$\mathcal{I}$-EDL	71.29 $\pm$ 0.0	84.14 $\pm$ 0.0	89.51 $\pm$ 2.5	42.87 $\pm$ 7.6	82.99 $\pm$ 2.3	86.81 $\pm$ 2.4
R-EDL	70.27 $\pm$ 0.1	82.45 $\pm$ 0.0	89.60 $\pm$ 2.6	45.13 $\pm$ 5.4	77.34 $\pm$ 3.6	89.40 $\pm$ 1.7
MoDEX	70.20 $\pm$ 0.1	82.56 $\pm$ 0.0	92.72 $\pm$ 1.3	87.92 $\pm$ 6.6	85.61 $\pm$ 3.6	49.32 $\pm$ 0.2
DMM	81.01 $\pm$ 0.3	70.71 $\pm$ 1.1	93.17 $\pm$ 1.2	87.93 $\pm$ 8.2	83.54 $\pm$ 1.1	87.14 $\pm$ 1.1
ETN	77.92 $\pm$ 0.0	78.29 $\pm$ 0.7	91.51 $\pm$ 1.2	54.81 $\pm$ 18.2	75.19 $\pm$ 3.6	64.00 $\pm$ 8.0
BEDI	86.44 $\pm$ 0.1	**86.67** $\pm$ 0.2	**93.79** $\pm$ 1.0	**91.06** $\pm$ 3.8	87.54 $\pm$ 2.5	**93.89** $\pm$ 5.3

Table 11: **(ID calibration: from-scratch setting)** In-distribution ECE / NLL over five seeds (mean $\pm$ standard deviation; lower is better). ECE uses 15 equal-width confidence bins.

Method	CIFAR-10	CIFAR-100	CIFAR-10-LT ($\rho = 0.1$)	CIFAR-10-LT ($\rho = 0.01$)
LLLA	0.047 $\pm$ 0.001 / 0.326 $\pm$ 0.008	0.019 $\pm$ 0.004 / 1.177 $\pm$ 0.015	0.091 $\pm$ 0.002 / 0.645 $\pm$ 0.013	0.226 $\pm$ 0.019 / 1.517 $\pm$ 0.110
VBLL	0.068 $\pm$ 0.002 / 0.488 $\pm$ 0.007	0.143 $\pm$ 0.003 / 1.373 $\pm$ 0.014	0.129 $\pm$ 0.004 / 0.920 $\pm$ 0.027	0.275 $\pm$ 0.025 / 1.976 $\pm$ 0.259
MoDEX	0.055 $\pm$ 0.004 / 0.357 $\pm$ 0.011	0.080 $\pm$ 0.005 / 1.312 $\pm$ 0.028	0.085 $\pm$ 0.024 / 0.658 $\pm$ 0.045	0.226 $\pm$ 0.015 / 1.588 $\pm$ 0.101
DMM	0.024 $\pm$ 0.003 / 0.242 $\pm$ 0.008	0.108 $\pm$ 0.001 / 1.053 $\pm$ 0.015	0.058 $\pm$ 0.004 / 0.460 $\pm$ 0.017	0.170 $\pm$ 0.008 / 1.119 $\pm$ 0.032
ETN	0.409 $\pm$ 0.027 / 0.792 $\pm$ 0.055	0.752 $\pm$ 0.004 / 4.262 $\pm$ 0.006	0.385 $\pm$ 0.040 / 0.873 $\pm$ 0.099	0.362 $\pm$ 0.029 / 1.362 $\pm$ 0.031
BEDI	**0.012** $\pm$ 0.001 / **0.227** $\pm$ 0.006	**0.012** $\pm$ 0.002 / **0.834** $\pm$ 0.012	**0.031** $\pm$ 0.004 / **0.407** $\pm$ 0.018	**0.133** $\pm$ 0.017 / **0.963** $\pm$ 0.036

Table 12: **(ID calibration: pretrained setting)** In-distribution ECE / NLL (mean $\pm$ standard deviation; lower is better). ECE uses 15 equal-width confidence bins. Results use five ImageNet seeds, fifteen Dry Bean seeds per held-out class, and three OpenBookQA seeds. A displayed standard deviation of 0.000 denotes a value below 0.0005.

Method	ImageNet	Dry Bean	OpenBookQA
LLLA	**0.022** $\pm$ 0.004 / **0.787** $\pm$ 0.003	**0.009** $\pm$ 0.003 / 0.130 $\pm$ 0.033	0.060 $\pm$ 0.015 / 0.403 $\pm$ 0.003
VBLL	0.029 $\pm$ 0.004 / 0.789 $\pm$ 0.002	0.012 $\pm$ 0.003 / 0.130 $\pm$ 0.033	0.040 $\pm$ 0.005 / 0.401 $\pm$ 0.040
MoDEX	0.518 $\pm$ 0.178 / 2.010 $\pm$ 0.850	0.014 $\pm$ 0.003 / 0.136 $\pm$ 0.035	0.100 $\pm$ 0.004 / 0.605 $\pm$ 0.065
DMM	0.509 $\pm$ 0.005 / 2.717 $\pm$ 0.030	0.010 $\pm$ 0.003 / 0.132 $\pm$ 0.033	0.097 $\pm$ 0.007 / 0.558 $\pm$ 0.028
ETN	0.799 $\pm$ 0.000 / 6.027 $\pm$ 0.005	0.233 $\pm$ 0.024 / 0.369 $\pm$ 0.038	0.615 $\pm$ 0.003 / 1.386 $\pm$ 0.000
BEDI	0.034 $\pm$ 0.004 / 0.792 $\pm$ 0.000	0.010 $\pm$ 0.004 / **0.124** $\pm$ 0.033	**0.038** $\pm$ 0.005 / **0.373** $\pm$ 0.016

the predictive-mean distortion identified in Proposition 4.1, this comparison highlights the value of preserving calibrated class probabilities during dispersion inference.

The calibration of BEDI is exactly that of its base softmax classifier after temperature scaling, since dispersion inference does not change the predictive mean. Thus, this calibration step yields competitive in-distribution probabilities while BEDI separately infers input-specific dispersion. Exploring other calibration methods, such as vector or matrix scaling, is beyond the scope of this work.

E.5 Ablation Study

Table 13 compares the full model with the frozen, calibrated base predictor and variants using either effective sample size or Bayesian predictive disagreement alone. The full model achieves the highest average OOD/shift AUPR across the five settings and outperforms both individual signals in four of them. On Dry Bean, disagreement alone performs slightly better (91.83 versus 91.71), indicating that the benefit of combining the signals varies across datasets. Overall, these results support the complementary roles of training support and Bayesian predictive disagreement in hierarchical dispersion inference.

Table 13: **(Ablation study)** Contribution of effective sample size (ESS) and Bayesian predictive disagreement to OOD/shift detection performance. Cells report OOD/shift AUPR averaged over the OOD sets of each setting (mean±std over seeds); *Base* is the frozen and calibrated predictor's max softmax probability, *ESS* reads the support estimate alone as the Dirichlet strength, and *Disagreement* scores the class-factorized last-layer posterior alone. Accuracy is identical across rows.

Method	CIFAR-10	CIFAR-10-LT ($\rho = 0.01$)	ImageNet	Dry Bean	OBQA	Avg.
Base	89.80 ± 0.4	74.30 ± 1.2	82.10 ± 0.0	89.55 ± 0.6	78.85 ± 1.1	82.92
ESS	86.50 ± 6.3	58.18 ± 13.8	84.59 ± 0.5	90.28 ± 0.7	72.12 ± 6.0	78.33
Disagreement	90.25 ± 0.4	74.84 ± 1.0	82.04 ± 0.0	$\mathbf{91.83} \pm 0.4$	79.48 ± 2.5	83.69
Full BEDI	$\mathbf{91.73} \pm 0.8$	$\mathbf{79.32} \pm 1.2$	$\mathbf{88.62} \pm 0.2$	91.71 ± 0.4	$\mathbf{91.78} \pm 7.2$	**88.63**

E.6 Controlled Study

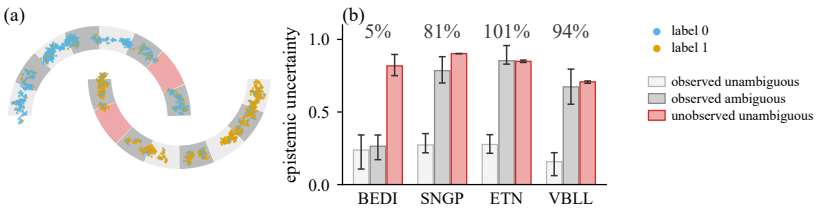

Figure 2: **Training support and label ambiguity in the controlled two-moons benchmark.** (a) Blue and orange points indicate the two labels. Light gray, dark gray, and pink backgrounds indicate observed unambiguous, observed ambiguous, and unobserved unambiguous regions, respectively. (b) Mean uncertainty score in each region, scaled separately for each method. Percentages report $100(\bar{u}_{\text{amb}} - \bar{u}_{\text{unamb}})/(\bar{u}_{\text{unobs}} - \bar{u}_{\text{unamb}})$, comparing the increase associated with observed ambiguity with that associated with missing support.

Benchmark construction. We use a two-moons benchmark to compare uncertainty from conflicting labels with uncertainty from missing training support. We divide each moon into eight equal angular cells. We sample 70 training anchors per moon, perturb their coordinates with isotropic Gaussian noise of standard deviation 0.035, and repeat observations at each resulting coordinate. Repetition counts follow $(20, 20, 8, 8, 20, 20, 8, 8)$ across the cells. Nominal label-flip rates alternate between 0.25 and 0.10; for an anchor with n observations and flip rate p, exactly $\text{round}(np)$ labels are flipped. Thus, every observed anchor has some label ambiguity, but its degree varies across regions. We exclude the segment $|\theta - 3\pi/16| < 0.22$ from training and all model-selection data to create an unobserved region with low label ambiguity.

Table 14: **(Limited training information vs. predictive ambiguity)** Quantitative evaluation on the controlled two-moons benchmark (mean $\pm$ standard deviation over three seeds). AUROC treats shallow or deep queries in the unobserved unambiguous region as positives and observed ambiguous queries as negatives. Distance ρ is the Spearman correlation between the uncertainty score and distance to the retained noiseless arcs; partial ρ controls for the frozen classifier's predictive entropy.

Method	Acc. (%)	Shallow unobserved AUROC	Deep unobserved AUROC	Distance ρ	Partial ρ
SNGP	100.0 ± 0.0	0.802 ± 0.075	0.857 ± 0.100	0.619 ± 0.089	0.588 ± 0.070
VBLL	100.0 ± 0.0	0.614 ± 0.097	0.620 ± 0.114	0.392 ± 0.100	0.288 ± 0.096
EDL	100.0 ± 0.0	0.575 ± 0.058	0.562 ± 0.061	0.110 ± 0.134	0.119 ± 0.066
DAEDL	99.2 ± 0.6	0.678 ± 0.071	0.790 ± 0.061	0.415 ± 0.065	0.391 ± 0.081
ETN	100.0 ± 0.0	0.427 ± 0.058	0.411 ± 0.073	0.208 ± 0.069	0.030 ± 0.108
BEDI	100.0 ± 0.0	$\mathbf{0.830 \pm 0.022}$	$\mathbf{0.958 \pm 0.016}$	$\mathbf{0.664 \pm 0.026}$	$\mathbf{0.630 \pm 0.012}$

Model fitting. We use separate validation, post-hoc fitting, and post-hoc selection splits containing 36, 42, and 42 anchors per moon, respectively. The base classifier for BEDI is a width-64 MLP with two residual blocks, trained for 600 epochs. We select it from three restarts by validation accuracy, breaking ties with validation NLL, and then freeze it. We select one of ten BEDI configurations using the marginal NLL of the disagreement-derived pseudo-observations on the post-hoc selection split and refit the selected configuration on both post-hoc splits. The removed segment is used only for evaluation.

Evaluation and results. We compare unobserved queries with observed ambiguous queries using each method's own uncertainty score. Shallow and deep unobserved queries have distances of $[0.025, 0.075]$ and $[0.155, 0.215]$, respectively, from the retained noiseless arcs. Each AUROC compares 1,800 unobserved queries with 2,400 observed ambiguous queries. We also measure how uncertainty varies with distance to the retained arcs on 8,000 diagnostic queries, reporting Spearman correlation and partial correlation controlling for the frozen classifier's predictive entropy. Classification accuracy is measured on 3,000 queries along the retained arcs. Results are averaged over three seeds.

In Figure 2, BEDI assigns similar uncertainty to the two observed regions but substantially higher uncertainty to the unobserved region. The increase associated with observed ambiguity is only 5% of the increase associated with missing support, compared with 81–101% for the displayed baselines. Table 14 shows the same pattern quantitatively: BEDI obtains the highest shallow and deep unobserved AUROCs (0.830 and 0.958), and its uncertainty remains strongly associated with distance after controlling for predictive entropy (partial $\rho = 0.630$).