VecUQ-OT: Aggregating Uncertainty Measures via Multivariate Ranks

Nikita Kotelevskii[1] **Vladimir Kondratyev**[1] **Maiya Goloburda**[1] **Alexander Fishkov**[1]
Mohsen Guizani[1] **Eric Moulines**[1,2] **Maxim Panov**[1]
[1]Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE
[2]LRE, EPITA, Paris, France

Abstract

Decision-making under uncertainty typically requires a single scalar score, forcing practitioners to commit to one uncertainty measure. However, different measures often capture complementary failure modes, and relying on a single measure may be insufficient for downstream tasks. We propose *VecUQ-OT*, a label-free procedure that aggregates multiple scalar uncertainty measures into a single ranking score via entropy-regularized optimal transport. Rather than selecting a single measure, our approach combines signals of (possibly) different natures through non-additive fusion based on *multivariate ranks*. The construction is motivated by the theory of multivariate ranks via exact optimal transport and realized through a tractable entropic approximation that generalizes to unseen inputs without retraining. Experiments across synthetic, image, and text domains demonstrate that *VecUQ-OT* provides stable performance across tasks, remains reliable when individual measures fail, and outperforms the natural additive fusion alternative.

1 Introduction

Uncertainty quantification (UQ) in machine learning (Hüllermeier and Waegeman, 2021) is a rapidly growing field, driven by the increasing deployment of artificial intelligence systems in critical applications (Begoli *et al.*, 2019; Kendall and Gal, 2017). For a given input, a predictive model typically only returns a prediction used in the downstream problem. Uncertainty quantification extends this output with a scalar score that reflects how much this prediction can be trusted.

A wide range of uncertainty measures have been proposed, including risk-based scores (Kotelevskii *et al.*, 2025; Schweighofer *et al.*, 2025), density-based estimators (Lee *et al.*, 2018), and ensemble-based methods (Lakshminarayanan *et al.*, 2017). However, no single uncertainty measure is universally robust across applications. Performance can vary significantly depending on the downstream task and dataset (Hofman *et al.*, 2026), forcing practitioners to carefully select or switch measures on a case-by-case basis.

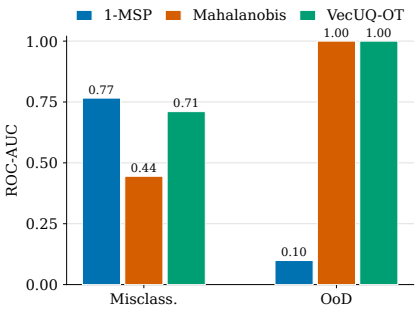


Figure 1: **Complementarity of uncertainty measures.** ROC-AUC on misclassification detection and OoD detection. 1-MSP and Mahalanobis each excel at only one task, while their fusion via VecUQ-OT is robust to both.

Ideally, one would select several reasonable complementary measures rather than committing to just one. But how can multiple measures be combined for the final decision? They may be of a different nature (e.g., risk-based vs. density-based) and operate on different scales. As a motivating example, consider two downstream problems: misclassification detection and out-of-distribution (OoD)

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

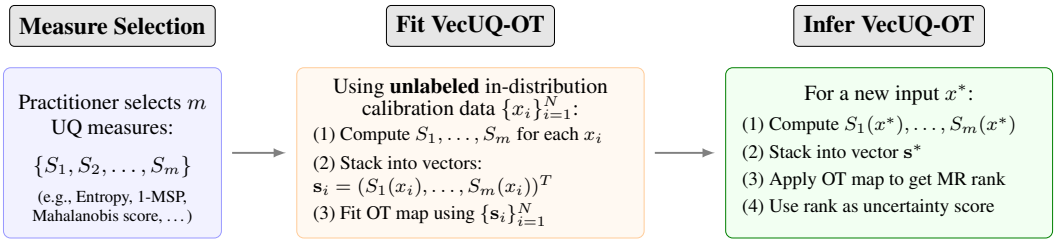


Figure 2: **VecUQ-OT workflow.** (Left) The practitioner selects a set of complementary uncertainty measures. (Middle) The OT is fit on unlabeled in-distribution data. (Right) For new inputs, VecUQ-OT produces a single uncertainty score via the learned mapping.

detection (see Figure 1 and Appendix A). We compare two uncertainty measures: 1-maximum softmax probability (1-MSP) (Hendrycks and Gimpel, 2017) and the Mahalanobis score (Lee *et al.*, 2018). Each measure performs well at only one task (see ROC-AUC values in Figure 1): 1-MSP detects misclassifications well but fails on OoD, while Mahalanobis shows the opposite pattern. Such complementarity is the norm rather than the exception. Even within a single task such as OoD detection, the best-performing measure often varies across datasets (Hofman *et al.*, 2026).

We propose a method that allows selecting multiple complementary uncertainty measures rather than committing to a single one. Our approach combines individual measures into a vector (for example, 1-MSP and the Mahalanobis score) and uses entropic-regularized optimal transport (OT) to generate a single, robust uncertainty score, that reflects the uncertainty, "contained" in the vector. We refer to the resulting method as **VecUQ-OT** *(Vector Uncertainty Quantification via OT)*. As seen in Figure 1, VecUQ-OT performs robustly for *both tasks simultaneously*, even when one component is weak.

The workflow of VecUQ-OT is as follows. The practitioner (1) selects a set of uncertainty measures that are individually reasonable for the task at hand, (2) fits an aggregation procedure on **unlabeled in-distribution** calibration data, and (3) for a new input x^*, computes all selected measures, stacks them into a vector, and obtains a single uncertainty score via the learned aggregation. We demonstrate it schematically in Figure 2.

The main **contributions** of this work can be summarized as follows.

1. We propose to aggregate several complementary uncertainty measures into a single score, rather than committing to one measure, and we cast this aggregation as a problem of ordering uncertainty vectors; see Section 2.
2. We construct a practical estimator of such an ordering. We build on the multivariate ranks of Chernozhukov *et al.* (2017); Hallin *et al.* (2021); Hallin and Konen (2024) and adapt them to uncertainty vectors. The method requires only unlabeled in-distribution data for fitting, and generalizes to unseen inputs without retraining; see Section 3.
3. We evaluate our approach across domains (synthetic, image, and text) and downstream UQ tasks (OoD detection, misclassification detection, selective prediction, and selective generation), demonstrating that VecUQ-OT provides consistently robust performance, even when individual measures fail; see Section 5.

2 Background

2.1 Vectorizing Uncertainty Signals

We start by introducing the setup and notation. We assume that the practitioner selects m uncertainty measures $S_1, \ldots, S_m$, where each measure $S_i \colon \mathbb{R}^d \to \mathbb{R}$ maps an input $x \in \mathbb{R}^d$ to a scalar uncertainty score. These measures may be of a completely different nature (e.g., risk-based (Kotelevskii *et al.*, 2025), density-based (Lee *et al.*, 2018)), but presumably each of them reasonably assesses the uncertainty for the task at hand. For an input object $x \in \mathbb{R}^d$, we evaluate all these measures and stack them into a vector:

$$\mathbf{s}(x) = \big(S_1(x), \ldots, S_m(x)\big)^\top \in \mathbb{R}^m.$$

Therefore, the vector contains different uncertainty signals at x.

For downstream problems that require decision-making based on uncertainty, we need to know how to compare these vectors. Namely, given two points x and x', we need to determine which of the vectors $\mathbf{s}(x)$ and $\mathbf{s}(x')$ reflects higher uncertainty. In other words, how to determine which vector is "bigger" (more uncertain) or "smaller" (less uncertain). For scalars, the answer is trivial, as there is a natural ordering on $\mathbb{R}$ that allows direct comparison.

For vectors, however, no canonical ordering exists. One can fix an order relation a priori, for example, by comparing norms or means of the coordinates, but such relations ignore the scales of the components and their dependencies. Instead, recent works (Chernozhukov *et al.*, 2017; Hallin *et al.*, 2021; Hallin and Konen, 2024) allow the data to determine the ordering of observations within a point cloud drawn from a common distribution. We follow this line of work and compare against distribution-agnostic alternatives in Section 5.

2.2 Multivariate Ranks

In the univariate case, the rank of an observation is its position in the sorted array. Interestingly, ranks can be characterized through an optimal transport formulation.

Univariate ranks via optimal transport. In dimension one, transporting the law of a scalar score S to the uniform distribution on $[0, 1]$ under quadratic cost recovers the cumulative distribution function F_S, and the inverse map is the quantile function (Santambrogio, 2015; Hallin *et al.*, 2021). For a single uncertainty measure, this is equivalent to replacing a raw score s by its quantile level $\tau = F_S(s)$. We want to retain this property for vectors of uncertainty measures. In dimension $m > 1$, no canonical cumulative distribution function exists, but one possible extension is provided by optimal transport (Chernozhukov *et al.*, 2017; Hallin *et al.*, 2021).

Multivariate ranks via optimal transport. Concretely, let $\mathbf{S}$ be a random uncertainty vector taking values in $\mathbb{R}^m$. We refer to its law ν as the *score distribution* and use *reference distribution* μ for the target of the transport map, both admitting a density with respect to Lebesgue measure. Assuming that both ν and μ have finite second moments and that ν gives no mass to $(m-1)$-dimensional surfaces, Brenier's theorem (Santambrogio, 2015) guarantees the existence of a unique (up to additive constants) convex function $\varphi\colon \mathbb{R}^m \to \mathbb{R}$ whose gradient defines a map $T \coloneqq \nabla\varphi$ that pushes ν forward to μ, minimizing the quadratic transportation cost:

$$T = \underset{\tilde{T}\colon \tilde{T}_{\#}\nu=\mu}{\arg\min} \int \|\tilde{T}(\mathbf{s}) - \mathbf{s}\|_2^2 \, d\nu(\mathbf{s}). \tag{1}$$

As the gradient of a convex function, the map T is *cyclically monotone* (Santambrogio, 2015), which is the natural multidimensional generalization of the univariate monotonicity property and provides the ordering structure exploited by multivariate ranks. When μ is chosen to be the uniform distribution on the unit ball, the resulting map induces a *center-outward ordering* (Chernozhukov *et al.*, 2017), and points mapped farther from the center of the reference distribution are considered more extreme. Intuitively, this ordering ranks multivariate observations according to their distance from the "median" of the distribution, identified with $T^{-1}(0)$.

Importantly, in our setting, a radially symmetric reference distribution is not appropriate. Uncertainty measures are typically scaled so that smaller values correspond to lower uncertainty, with the minimum attained at their origin. To preserve this interpretation, we choose a reference measure μ supported on a non-negative subspace, so that the least uncertain observations are mapped near the origin and thus receive the lowest ranks (see Figure 6 in Appendix A for the illustration of the resulting ordering). Reference distributions that are not radially symmetric and induce a similar left-to-right ordering were also used to measure multivariate risk (Bercu *et al.*, 2024) and multidimensional inequality (Fan *et al.*, 2024). In analogy to Monge-Kantorovich rank (Chernozhukov *et al.*, 2017), we define the *scalar multivariate rank (MR)* as

$$r(\mathbf{s}) = \|T(\mathbf{s})\|_2, \qquad \mathbf{s} \in \mathbb{R}^m.$$

In the exact OT setting, these ranks satisfy several useful properties (Chernozhukov *et al.*, 2017; Hallin *et al.*, 2021; Hallin and Konen, 2024), including Glivenko-Cantelli consistency of empirical ranks, distribution-freeness of the resulting rank statistics under absolutely continuous score distributions, and recovery of the natural ordering on $\mathbb{R}$ in the univariate case.

3 VecUQ-OT

We now apply the multivariate rank to order **s**core vectors $\mathbf{s}(x)$, composed of uncertainty measures. Let $\mathcal{D}_{\text{cal}} = \{x_i\}_{i=1}^N$ be a calibration set of *unlabeled*, in-distribution inputs, and let $\mathbf{s}_i = \mathbf{s}(x_i) \in \mathbb{R}^m$ denote the corresponding uncertainty vectors. Let $\hat{\nu}$ denote the empirical distribution of $\{\mathbf{s}_i\}_{i=1}^N$, and let μ be a reference distribution on $\mathbb{R}^m$ (see Appendix B for specific choices). We discretize μ into $\hat{\mu}$ via N target points $\{\tilde{\mathbf{s}}_j\}_{j=1}^N$ with associated masses $\{\hat{\mu}_j\}_{j=1}^N$.

Following Section 2.2, we seek a transport map $T\colon \mathbb{R}^m \to \mathbb{R}^m$ given by equation (1), where a sum over the empirical support replaces the integral. For any input x, we define its *scalar rank* as $r(x) = \|T(\mathbf{s}(x))\|_2$. We say $\mathbf{s}(x)$ is *more uncertain* than $\mathbf{s}(x')$ if $r(x) > r(x')$.

This construction is well-suited to aggregating heterogeneous uncertainty measures for three reasons. First, fitting the rank requires only unlabeled in-distribution covariates, since labels do not enter the OT problem. Second, the rank map T is a nonlinear transformation that accounts for differences in scale and cross-correlations across components, which is essential when the input measures arise from distinct frameworks (e.g., risk-based and density-based) and operate on incomparable scales. Third, unlike fixed rules such as the sum of standardized components, the ordering depends on the joint distribution of the components, making it much less sensitive to redundant components (Section 5.4).

From exact OT to a practical estimator. The construction in Section 2.2 relies on the exact OT map between $\hat{\nu}$ and $\hat{\mu}$. While exact OT provides the *theoretical motivation* for our ranking, the exact discrete OT solution is impractical in our setting for two reasons. First, discrete OT is an assignment problem between fixed sets of source and target points, so evaluating the rank of a new test point would require re-solving an assignment problem from scratch. Second, two distinct test points yield two distinct exact OT solutions, and ranks computed from different solutions are not directly comparable. Both issues prevent natural out-of-sample evaluation of new test-time uncertainty vectors, which is essential for our use case.

To make the method practically applicable, we therefore use entropy-regularized OT (Cuturi, 2013) combined with barycentric projection, detailed below. This estimator is fit only once on the calibration set and extends to any new input via a closed-form weighted average, without recomputing the transport plan. Exact OT is recovered in the limit $\epsilon \to 0$, in which the theoretical guarantees of Section 2.2 hold exactly. For $\epsilon > 0$, those guarantees no longer hold strictly, but for sufficiently small ϵ the induced ordering retains the same qualitative behavior. We verify this empirically in Appendix H, where the ordering produced by VecUQ-OT matches the exact OT case in scenarios where the latter is unambiguously defined (single-component vectors, duplicated coordinates, and constant-coordinate signals).

Entropic OT solution. Let $C \in \mathbb{R}^{N\times N}$ be the cost matrix, with elements $C_{ij} = \|\mathbf{s}_i - \tilde{\mathbf{s}}_j\|_2^2$. We solve the following regularized problem:

$$\min_{P\in\Pi(\hat{\nu},\hat{\mu})} \langle C, P\rangle + \epsilon \sum_{i,j} P_{ij}\log P_{ij},$$

where $\Pi(\hat{\nu},\hat{\mu}) = \{P \geq 0\colon P\mathbf{1} = \hat{\nu},\ P^\top\mathbf{1} = \hat{\mu}\}$ is the set of valid transport plans, and $\epsilon > 0$ controls the strength of entropic regularization. The solution has the Sinkhorn form (Cuturi, 2013):

$$P^\epsilon = \operatorname{diag}(u) K \operatorname{diag}(v), \quad K_{ij} = \exp(-C_{ij}/\epsilon),$$

with scaling vectors $u, v \in \mathbb{R}^N$ chosen to satisfy the marginal constraints.

Barycentric projection for out-of-sample points. The transport plan P^ϵ is defined only on the calibration points $\{\mathbf{s}_i\}$. To evaluate the rank of a new input x with uncertainty vector $\mathbf{s} = \mathbf{s}(x)$, we use the *barycentric projection* (Peyré *et al.*, 2019; Pooladian and Niles-Weed, 2021):

$$\widehat{R}_\epsilon(x) = \sum_{j=1}^N \tilde{\mathbf{s}}_j w_j^\epsilon(\mathbf{s}), \tag{2}$$

where the weights are given by

$$w_j^\epsilon(\mathbf{s}) = \frac{v_j \exp\{-\|\mathbf{s} - \tilde{\mathbf{s}}_j\|_2^2/\epsilon\}}{\sum_{\ell=1}^N v_\ell \exp\{-\|\mathbf{s} - \tilde{\mathbf{s}}_\ell\|_2^2/\epsilon\}}.$$

Intuitively, $\widehat{R}_\epsilon(x)$ computes a soft assignment of $\mathbf{s}$ to the reference points. By construction, $\widehat{R}_\epsilon(x)$ is a weighted average of the reference points, so it always lies within their convex hull. As a result, the ranks of OoD inputs, whose uncertainty vectors typically have large components, saturate. However, what matters for downstream tasks is the order of the ranks rather than their values, and Appendix B.1 shows that OoD inputs are still ranked above in-distribution inputs. The final aggregated uncertainty score is the scalar rank:

$$r(x) = \|\widehat{R}_\epsilon(x)\|_2,$$

with higher values indicating more uncertainty in the inputs. We refer to this score as **VecUQ-OT**.

Discussion of VecUQ-OT. The multivariate rank discards the scale information present in the original uncertainty measures (e.g., variance, which is proportional to the expected error), which is essential for interpretability. However, most downstream applications of uncertainty quantification depend only on the *ordering* induced by the scores, not their absolute values. For instance, OoD detection is typically evaluated via ROC-AUC, which is purely rank-based. Similarly, selective prediction operates by thresholding, where inputs exceeding a chosen uncertainty level are excluded, and only the relative ordering determines which inputs are deferred. The same reasoning applies to other uncertainty-driven tasks.

Beyond capturing complementary failure modes, aggregation offers another benefit. Each individual uncertainty measure may capture a meaningful signal related to uncertainty, such as the predictive distribution's variance. However, we usually deal only with empirical estimates that are imperfect and contain errors relative to the ground-truth quantity. By aggregating multiple measures, even when they are of different natures, we can potentially reduce estimation noise. The resulting score may lack direct physical interpretation, but it provides a more robust scalar uncertainty signal than any individual component.

In practice, the choice of components is left to the practitioner and follows the same domain expertise that already drives the selection of an individual uncertainty measure for a given task. VecUQ-OT does not aim to replace this step, nor to dominate every component on every individual benchmark. Rather, it reduces the risk of committing to a single measure by combining several reasonable signals into one score, which remains reliable even when some components fail (see Figure 1). As a practical default, we recommend selecting a small set of complementary measures that capture distinct sources of uncertainty (e.g., risk-based and density-based), rather than several variants of the same family.

The values produced by VecUQ-OT can be interpreted as *soft ranks* that order uncertainty vectors according to how typical or extreme they are. Our empirical evaluation (see Section 5) shows that the ordering they induce is meaningful and yields robust performance across diverse datasets and tasks.

Computational cost. Fitting VecUQ-OT reduces to solving a single discrete entropic OT problem on the calibration set via Sinkhorn iterations (Cuturi, 2013), a convex problem with efficient off-the-shelf solvers and no additional parametric model to train. The dominant cost is constructing and operating on the $N \times N$ kernel matrix, giving an overall fitting complexity of $\mathcal{O}\big(N^2(m+K)\big)$, where K is the number of Sinkhorn iterations. Inference for a new input is a closed-form weighted average over the calibration support and runs in $\mathcal{O}(Nm)$ per query. Actual run time measurements for different choices of N and m are reported in Appendix M.

4 Related Work

Uncertainty quantification. A large body of work in UQ proposes measures targeting different types of uncertainty (Hüllermeier and Waegeman, 2021). The most common distinction is between *aleatoric uncertainty* (AU), which arises from inherent randomness between inputs and outputs and cannot be reduced by collecting more data, and *epistemic uncertainty* (EU), which reflects the lack of knowledge about the data-generating process and can be reduced as more data become available. For AU, any notion of dispersion (e.g., entropy, variance) of the true data-generating distribution $p(y \mid x)$ can serve as a measure (Kotelevskii *et al.*, 2025; Schweighofer *et al.*, 2025; Hofman *et al.*, 2026; Fishkov *et al.*, 2026). For the EU, the landscape is more diverse, with measures ranging from prediction disagreement in ensembles or Bayesian models (Lakshminarayanan *et al.*, 2017; Gal and Ghahramani, 2016) to density-based approaches that use learned representations (Lee *et al.*, 2018; Kotelevskii *et al.*, 2022).

In practice, a practitioner typically commits to a single measure, even though different measures (e.g., risk-based and density-based) operate on incomparable scales and arise from distinct frameworks, making direct comparison or combination difficult.

Beyond the AU/EU split, a common quantity in the UQ literature is *total uncertainty* (TU), used when overall reliability matters more than separating its sources. TU is typically obtained by adding AU and EU, which arises naturally in risk-based frameworks (Kotelevskii *et al.*, 2025), but has been criticized for failing to hold in practice (Wimmer *et al.*, 2023). From an aggregation perspective, additive TU is one specific recipe for combining two related signals, and that recipe is unlikely to extend cleanly to arbitrary heterogeneous measures. VecUQ-OT addresses both issues by aggregating a set of measures into a single score.

Optimal transport and multivariate ordering. OT provides a geometric framework for comparing probability distributions by finding a minimal-cost mapping between them (Peyré *et al.*, 2019). Entropic regularization enables scalable computation via Sinkhorn iterations (Cuturi, 2013). Beyond distribution matching, OT has been connected to multivariate quantiles and ranks (Chernozhukov *et al.*, 2017; Hallin *et al.*, 2021; Hallin and Konen, 2024), providing a notion of vector ordering. Smooth and regularized estimators of these transport maps have also been proposed (Hallin *et al.*, 2021; Beirlant *et al.*, 2020; Bercu *et al.*, 2026). These constructions have enabled practical methods, including multidimensional conformal prediction (Thurin *et al.*, 2025; Klein *et al.*, 2026) and multivariate nonparametric testing (Ghosal and Sen, 2022). Closest in spirit to our goal, OT-based quantiles and ranks have been used to measure multivariate risks (Ekeland *et al.*, 2012; Beirlant *et al.*, 2020; Bercu *et al.*, 2024). We build on these works to define an ordering over uncertainty vectors.

Relation to OT-based UQ methods. OT-based multivariate ranks have recently been used in two adjacent lines of work. Vector quantile regression (Carlier *et al.*, 2016) fits an OT map to model the conditional distribution of a target Y given covariates Z, and therefore requires labeled data for loss optimization. OT-based conformal prediction (Thurin *et al.*, 2025) uses the Monge-Kantorovich rank map to assign nonconformity scores to a fixed grid of reference vectors, producing prediction sets with coverage guarantees. In that construction, each score is mapped to exactly one reference vector, so the rank takes only discrete values from the predefined grid.

VecUQ-OT differs from both in three ways. First, it is *fully unsupervised* and uses only unlabeled in-distribution covariates for calibration. Second, it uses barycentric projection rather than hard assignment, yielding soft, continuous ranks and a more flexible ordering. Third, the goal is different. We do not model a conditional distribution or construct prediction sets, but rather produce a scalar uncertainty score that orders inputs for downstream tasks such as OoD detection, misclassification detection, and selective prediction. More broadly, conformal prediction (Vovk *et al.*, 2005) produces prediction sets with coverage guarantees from nonconformity scores, whereas VecUQ-OT produces a scalar uncertainty score and makes no coverage claims.

5 Experiments

This section evaluates our approach across tasks, datasets, and modalities. As detailed in Appendix E, in all experiments we assume that no explicit OoD data are available during training, which is a realistic scenario in practice. A smaller subset (1.8k samples for all datasets) is used to fit the OT map and is referred to as the calibration set, while the remainder is used for evaluation. Note that only covariates are required for calibration, and no labels are needed. See Appendix F for an ablation on calibration set size. Unless stated otherwise, VecUQ-OT uses a uniform reference distribution on $[0, 1]^m$ with $\epsilon = 0.5$ (see Appendix C for other choices). Additional experiments, including analysis of extreme composition scenarios, are provided in Appendix H.

Evaluation protocol. We evaluate our aggregation method, along with other alternatives. For each method, the resulting aggregation produces a scalar score for each input, and we compare methods only using ranking-based metrics, such as ROC-AUC or PRR, without selecting a particular threshold. For OoD detection, the ROC-AUC is computed between the scores of in-distribution test inputs (negatives) and OoD inputs (positives). For misclassification detection, the positives are the test inputs on which the prediction is wrong. For selective prediction, we sort test inputs by increasing

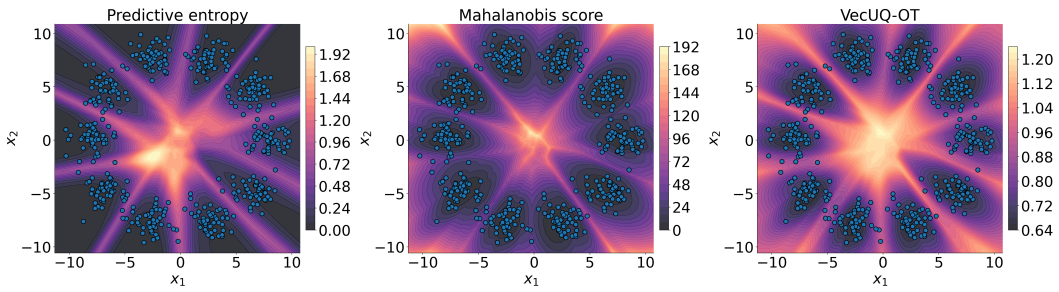


Figure 3: Uncertainty measures on the synthetic dataset. *Left:* predictive entropy (AU). *Middle:* Mahalanobis score (EU proxy). *Right:* VecUQ-OT, which aggregates both signals into a single score.

uncertainty, compute the accuracy on the first k inputs for each coverage level k/n, and report the area under the resulting accuracy-coverage curve.

Additive baseline. We compare VecUQ-OT against an additive aggregation baseline that takes the raw sum of components, $U_{\text{add}}(x) = \sum_{j=1}^{m} S_j(x)$, without learned weights or normalization. In the particular case of two components, where one estimates AU and another EU, this recovers the classical additive TU.

5.1 Synthetic Experiments

We begin with a two-dimensional synthetic classification problem with ten classes, where each class is a Gaussian blob (standard deviation 1) uniformly placed on a circle of radius 8 (see Figure 3). As a model, we use a single deterministic neural network with one hidden layer of 32 ReLU units.

Our goal is to quantify the predictive uncertainty of this model. We consider two uncertainty measures, one capturing AU and the other EU. AU is estimated via predictive entropy, a standard risk-based measure. Since there is no ensemble or Bayesian model to measure EU directly, we use the Mahalanobis distance as a proxy. The components are computed as follows:

$$\mathrm{AU}(x) = H\big[p_\theta(y \mid x)\big] = -\sum_{c=1}^{C} p_c \log p_c,$$

where $p_c = p_\theta(y = c \mid x)$, and

$$\mathrm{EU}(x) = M(x) = \min_{c=1,\dots,C} \|f_\theta(x) - \hat{\mu}_c\|^2_{\hat{\Sigma}^{-1}},$$

where $\hat{\mu}_c = \frac{1}{N_c} \sum_{i:y_i=c} f_\theta(x_i)$ and $\hat{\Sigma} = \frac{1}{N} \sum_{c=1}^{C} \sum_{i:y_i=c} \big(f_\theta(x_i) - \hat{\mu}_c\big)\big(f_\theta(x_i) - \hat{\mu}_c\big)^\top$.

Given these measures, we form a two-dimensional uncertainty vector $(\mathrm{AU}, \mathrm{EU})$. Interestingly, in this particular case, since the only sources within the vector are AU and EU, the aggregated vector can be informally thought of as the "total" uncertainty. Applying VecUQ-OT to these vectors, we obtain a map shown in the rightmost panel of Figure 3. Regions of high VecUQ-OT align with areas of label ambiguity (where predictive entropy peaks, left panel) and sparse coverage (where the Mahalanobis score is large, middle panel). We see that in this particular case, the aggregated score behaves as expected for "total" uncertainty. The experiment demonstrates that our method effectively fuses the two sources of uncertainty into a single score that highlights both class overlap and OoD regions.

5.2 Image Datasets

For image data, we study three downstream UQ tasks: *OoD detection*, *misclassification detection*, and *selective prediction*. As base classifiers, we train four independent deep ensemble groups of ResNet18 (He *et al.*, 2016), each consisting of five members (20 models total), trained using different random seeds. For risk-based uncertainties, we consider Logscore instantiations (Kotelevskii *et al.*, 2025).

In-Dist.	Out-of-Dist.	Ours	Additive	Individual Components			
				$R^{1,2}_{\text{exc}}$	$R^{1,2}_{\text{tot}}$	R^{1}_{b}	MahS
CIFAR10	CIFAR100	**0.919 ± 0.001**	0.913 ± 0.001	0.907 ± 0.000	0.916 ± 0.001	0.917 ± 0.001	0.912 ± 0.002
CIFAR10	SVHN	0.959 ± 0.003	0.937 ± 0.005	0.946 ± 0.011	**0.963 ± 0.003**	**0.963 ± 0.002**	0.934 ± 0.005
CIFAR10	TinyImageNet	**0.914 ± 0.001**	0.911 ± 0.001	0.899 ± 0.000	0.910 ± 0.001	0.911 ± 0.001	0.910 ± 0.001
CIFAR100	CIFAR10	0.767 ± 0.001	0.552 ± 0.004	0.737 ± 0.001	**0.775 ± 0.002**	0.773 ± 0.002	0.535 ± 0.004
CIFAR100	SVHN	**0.873 ± 0.005**	0.701 ± 0.032	0.776 ± 0.012	0.870 ± 0.006	0.870 ± 0.007	0.679 ± 0.034
CIFAR100	TinyImageNet	**1.000 ± 0.000**	0.654 ± 0.005	**1.000 ± 0.000**	0.926 ± 0.001	0.810 ± 0.001	0.623 ± 0.005
TinyImageNet	ImageNet-A	**0.847 ± 0.002**	0.454 ± 0.009	0.801 ± 0.004	0.841 ± 0.002	0.835 ± 0.002	0.441 ± 0.009
TinyImageNet	ImageNet-R	**0.834 ± 0.002**	0.417 ± 0.008	0.793 ± 0.003	0.831 ± 0.003	0.825 ± 0.003	0.405 ± 0.008
TinyImageNet	ImageNet-O	0.752 ± 0.003	0.520 ± 0.005	**0.755 ± 0.002**	0.735 ± 0.003	0.724 ± 0.004	0.513 ± 0.005
Avg		**0.874 ± 0.002**	0.673 ± 0.008	0.846 ± 0.004	0.863 ± 0.002	0.848 ± 0.003	0.661 ± 0.008

Table 1: ROC-AUC of OoD detection. Best results are shown in bold and second-best results are underlined per row. The Logscore proper scoring rule instantiates risk-based measures (notation follows Kotelevskii *et al.* (2025)). MahS denotes the Mahalanobis score. "Ours" refers to VecUQ-OT. Additive denotes an alternative aggregation strategy, while the remaining methods report individual uncertainty components.

In-Dist.	Ours	Additive	Individual Components			
			$R^{1,2}_{\text{exc}}$	$R^{1,2}_{\text{tot}}$	R^{1}_{b}	MahS
CIFAR10	0.944 ±.001	0.930 ±.002	0.942 ±.003	**0.945** ±.002	0.942 ±.002	0.928 ±.003
CIFAR100	**0.851** ±.003	0.595 ±.006	0.832 ±.002	0.850 ±.003	0.845 ±.003	0.574 ±.006
TinyImageNet	0.849 ±.002	0.429 ±.004	0.831 ±.001	**0.850** ±.003	0.845 ±.003	0.417 ±.004
Avg	**0.882** ±.002	0.651 ±.004	0.868 ±.002	**0.882** ±.002	0.877 ±.002	0.639 ±.004

In-Dist.	Ours	Additive	Individual Components			
			$R^{1,2}_{\text{exc}}$	$R^{1,2}_{\text{tot}}$	R^{1}_{b}	MahS
CIFAR10	**0.997** ±.000	0.996 ±.000	**0.997** ±.000	**0.997** ±.000	**0.997** ±.000	0.996 ±.000
CIFAR100	0.916 ±.001	0.820 ±.001	0.913 ±.001	**0.918** ±.001	0.916 ±.001	0.811 ±.001
TinyImageNet	0.887 ±.001	0.670 ±.005	0.885 ±.001	**0.891** ±.000	0.889 ±.001	0.660 ±.005
Avg	0.933 ±.001	0.829 ±.002	0.932 ±.001	**0.935** ±.000	0.934 ±.001	0.822 ±.002

Table 2: Aggregate evaluation across uncertainty quantification tasks. The left table reports ROC-AUC for misclassification detection, while the right table reports the area under the accuracy-coverage curve for selective prediction. Best results are shown in bold and second-best results are underlined per row. "Ours" refers to VecUQ-OT. Additive denotes an alternative aggregation strategy, and the remaining methods report individual uncertainty components.

For all problems we consider, we use four components to form a vector: the specific total, excess, and Bayes risk approximations (following Kotelevskii *et al.* (2025)) and the Mahalanobis score. Other combinations are reported in Appendix I. Specifically, we consider risk-based measures that lead to the widely used expected entropy/conditional mutual information decomposition (Houlsby *et al.*, 2011; Kendall and Gal, 2017). We enhance these risk-based measures by adding another signal of a different nature, the Mahalanobis score. All our compositions mix signals of different natures, namely 1-MSP and the Mahalanobis score in Figure 1, risk-based measures and MahS here, and sequence-level CoCoA scores and token-level MSP in Section 5.3.

OoD detection. We train models on CIFAR10, CIFAR100 (Krizhevsky, 2009), and TinyImageNet (Le and Yang, 2015). For CIFAR10/100, OoD datasets are CIFAR10, CIFAR100, TinyImageNet, and SVHN (Netzer *et al.*, 2011). For TinyImageNet, OoD datasets are ImageNet-A, ImageNet-O (Hendrycks *et al.*, 2021c), and ImageNet-R (Hendrycks *et al.*, 2021a). We report ROC-AUC for ID vs. OoD ranking (see Table 1). Across datasets, VecUQ-OT remains consistently competitive, it (i) avoids the severe failures observed for some components, (ii) often is or close to the strongest component, and (iii) achieves the best average performance.

This experiment illustrates the practical value of VecUQ-OT. It does not always outperform the best individual component, but its performance remains stable when some components fail. For example, MahS performs poorly when CIFAR100 or TinyImageNet is the in-distribution dataset, whereas VecUQ-OT remains close to the best-performing component in these cases. Moreover, the best component changes from one dataset pair to another. Therefore, VecUQ-OT is especially useful when the best component for a given task is not known in advance (see Appendix L).

Misclassification detection. We use CIFAR10, CIFAR100, and TinyImageNet as IDs and report ROC-AUC for separating misclassified and correctly classified samples. Results are shown in Table 2 (left). VecUQ-OT is consistently robust, reaching the best or second-best score in all cases.

	LLaMA-8B			Mistral-7B			Falcon-7B			
Method	Trivia	WMT19DeEn	MMLU	Trivia	WMT19DeEn	MMLU	Trivia	WMT19DeEn	MMLU	Mean
Ours (Uniform, FW)	0.600	**0.614**	0.482	**0.680**	**0.643**	0.472	**0.699**	**0.615**	0.543	**0.594**
Ours (Exp, FW)	0.597	<u>0.612</u>	0.485	<u>0.678</u>	<u>0.641</u>	0.473	<u>0.698</u>	<u>0.612</u>	0.543	<u>0.593</u>
Additive	0.590	0.510	0.491	0.671	0.525	<u>0.474</u>	**0.699**	0.486	<u>0.545</u>	0.554
CoCoA MSP	0.603	0.584	<u>0.492</u>	0.677	0.607	0.469	**0.699**	0.590	0.539	0.584
CoCoA PPL	0.598	0.509	0.458	<u>0.678</u>	0.571	0.469	0.683	0.573	0.539	0.564
CoCoA NMTE	0.604	0.505	0.408	0.676	0.565	0.449	0.689	0.568	0.528	0.555
MSP	0.538	0.469	**0.516**	0.634	0.473	**0.478**	0.680	0.420	**0.548**	0.528
PPL	0.520	0.402	0.469	0.636	0.484	**0.478**	0.653	0.520	**0.548**	0.523
Consistency	<u>0.615</u>	0.450	0.395	0.652	0.499	0.427	0.657	0.487	0.493	0.520
SAR	0.593	0.472	0.360	0.638	0.519	0.418	0.647	0.503	0.520	0.519
MTE	0.506	0.391	0.362	0.623	0.477	0.459	0.642	0.532	0.543	0.504
DegMat	**0.616**	0.357	0.346	0.651	0.385	0.410	0.663	0.440	0.494	0.485
EigValLaplacian	0.600	0.283	0.296	0.622	0.330	0.401	0.656	0.400	0.471	0.451
Semantic Entropy	0.541	0.410	0.235	0.565	0.409	0.387	0.593	0.411	0.474	0.447
Mean	0.580	0.469	0.414	0.649	0.509	0.447	0.668	0.511	0.523	0.530

Table 3: Prediction Rejection Ratio (PRR; higher is better) for selective generation. Columns group three models (LLaMA-8B, Mistral-7B, Falcon-7B) and three datasets each (Trivia, WMT19DeEn, MMLU); the last column reports the mean across all nine entries. The top block shows our OT-based combinations (*VecUQ-OT*); the second block reports an additive aggregation baseline. The remaining blocks report individual components and reference baselines. **Best** and <u>second-best</u> per column are highlighted.

Selective prediction. For this problem, the metric is the *area under the accuracy-coverage curve*. Samples are sorted by uncertainty in ascending order, added incrementally, and accuracy is computed at each coverage level. The area under the resulting curve is the score (the higher, the better). Results in Table 2 (right) show that VecUQ-OT is again robust across datasets.

5.3 Text Datasets

We also evaluate on text data, focusing on *selective generation*, using uncertainty as a proxy for quality and rejecting outputs accordingly. Dataset and baseline details appear in Appendix J.

We use the Prediction Rejection Ratio (PRR) (Malinin and Gales, 2021) as the performance metric. PRR measures how average output quality improves as an increasing fraction of outputs is rejected based on uncertainty. Formally, PRR is the ratio of the area between the Prediction-Rejection curve for a given uncertainty score and a random baseline, to the area between the ideal oracle (which perfectly ranks instances by quality) and the random baseline:

$$\mathrm{PRR} = \frac{\mathrm{AUC}_{\mathrm{unc}} - \mathrm{AUC}_{\mathrm{rnd}}}{\mathrm{AUC}_{\mathrm{oracle}} - \mathrm{AUC}_{\mathrm{rnd}}}.$$

As individual uncertainty measures, we use those in the middle block of Table 3 (the CoCoA variants with MSP). Table 3 summarizes results for the selected datasets, and additional results are provided in Appendix J. The results show that our method provides robust aggregation across datasets and models. We also consider a composition of signals of different types (probabilistic, entropy-based, and semantic) in Table 21 of Appendix J. There, VecUQ-OT again gives the best average PRR among label-free aggregation methods, and only a model trained with labels performs better. We also report the Pareto-front hit rate of VecUQ-OT (Ours) in Figure 4, using the same individual components as in Table 3, and again the aggregation provided by our method is more robust.

5.4 Comparison Across Domains

Comparison with other aggregation rules. A vector of uncertainty measures can also be reduced to a scalar by simple heuristic rules, such as the sum of standardized components or the average of per-component ranks (Appendix L). To our knowledge, such rules have no theoretical justification as orderings of uncertainty vectors, and we do not provide one here. VecUQ-OT, in contrast, is based on the theory of multivariate ranks (Section 2.2). Still, the heuristic rules are natural alternatives worth comparing with. Table 4 shows that, on average, several of them perform close to VecUQ-OT. As a reference, we also include a *supervised model* trained on labels from the target task. It performs slightly better than all label-free methods, as expected, since it uses labels and therefore has more

Aggregation	OoD	Miscls.	Sel. pred.	PRR
VecUQ-OT (ours)	$\underline{0.8740 \pm 0.0020}$	$\underline{0.8815 \pm 0.0021}$	$\underline{0.9335 \pm 0.0006}$	$\underline{0.5942 \pm 0.0064}$
Additive (raw sum)	0.6733 ± 0.0078	0.6510 ± 0.0041	0.8287 ± 0.0021	0.5543 ± 0.0069
Standardized sum	0.8721 ± 0.0021	0.8736 ± 0.0022	0.9294 ± 0.0005	0.5934 ± 0.0063
Mean of component ranks	0.8615 ± 0.0028	0.8757 ± 0.0026	0.9323 ± 0.0006	0.5912 ± 0.0062
L2 norm after min-max	0.8713 ± 0.0024	0.8765 ± 0.0022	0.9307 ± 0.0009	0.5906 ± 0.0064
Whitened sum	0.8070 ± 0.0070	0.8006 ± 0.0098	0.9119 ± 0.0019	0.5649 ± 0.0066
Supervised model (uses labels)	$\mathbf{0.8773 \pm 0.0018}$	$\mathbf{0.8833 \pm 0.0022}$	$\mathbf{0.9352 \pm 0.0005}$	$\mathbf{0.6016 \pm 0.0063}$

Table 4: Comparison of alternative aggregation rules for uncertainty vectors. All rules aggregate the same uncertainty measures as in Tables 1–3 and are evaluated with the same metrics, averaged over datasets (and over LLMs for PRR). See Appendix L for details. The std is over ensemble groups for images and over 1000 bootstrap resamples for PRR. Best in bold, second-best underlined.

Method	Pareto front (%)
Ours	**84.76**
$R^{1,2}_{\text{tot}}$	$\underline{65.71}$
$R^{1,2}_{\text{exc}}$	25.71
R^{1}_{b}	5.71
Additive	0.0
MahS	0.0

Table 5: Aggregate Pareto-front hit rate across all task pairs. Best in bold, second-best underlined.

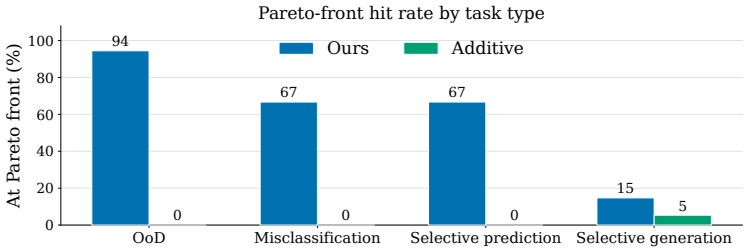


Figure 4: Per-task-type Pareto-front hit rate for the aggregation methods. The front for each pair is computed over both methods and the individual components.

information. However, the heuristic rules have failure modes. For example, they give a duplicated measure a larger weight, so when we add copies of MahS to the vector, the standardized sum and the mean of component ranks degrade much more than VecUQ-OT (Table 25). Appendix L describes all aggregation rules in detail, including how the supervised model is trained and which labels it uses.

Pareto analysis across tasks. Our goal is robust performance across tasks rather than the highest score on any single dataset. To summarize performance across task pairs, we count how often a method lies on the Pareto front defined by the two corresponding metrics (allowing pairs of the same task type across different datasets). Table 5 shows that our method most often appears on the Pareto front, indicating strong overall robustness. In Figure 4, we provide a detailed split by task and the number of appearances (hit rate) on the Pareto front. For OoD, misclassification, and selective prediction, we use the same combination of individual scores as discussed above.

6 Conclusion

We introduced VecUQ-OT, a method for ordering uncertainty vectors via multivariate ranks obtained from entropy-regularized OT. Multivariate ranks have been used to summarize multivariate risks in other settings (Ekeland *et al.*, 2012; Beirlant *et al.*, 2020; Bercu *et al.*, 2024). To the best of our knowledge, ours is the first work to apply them to the aggregation of predictive uncertainty measures. Alternative instantiations are possible, such as amortized OT (Amos, 2023) or normalizing flows with cyclical-monotonicity constraints (Huang *et al.*, 2021). However, they require training additional parametric models, introducing parameter-estimation uncertainty. VecUQ-OT is nonparametric and requires no model training beyond solving a convex optimization problem.

Empirically, VecUQ-OT provides a robust, label-free method for aggregating heterogeneous uncertainty measures. It performs competitively across domains and tasks without requiring commitment to a single scalar measure *a priori*. Unlike alternative heuristic aggregation rules, it does not degrade strongly when components are duplicated. Moreover, it needs only unlabeled in-distribution data for calibration and no additional labels, which are often expensive to obtain.

References

Brandon Amos. On amortizing convex conjugates for optimal transport. In *ICLR*, 2023.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, et al. Findings of the 2019 conference on machine translation. In *WMT*, 2019.

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 2019.

Jan Beirlant, Sven Buitendag, Eustasio Del Barrio, Marc Hallin, and Francois Kamper. Center-outward quantiles and the measurement of multivariate risk. *Insurance: Mathematics and Economics*, 95:79–100, 2020.

Bernard Bercu, Jérémie Bigot, and Gauthier Thurin. Monge-Kantorovich superquantiles and expected shortfalls with applications to multivariate risk measurements. *Electronic Journal of Statistics*, 18(2):3461–3496, 2024.

Bernard Bercu, Jérémie Bigot, and Gauthier Thurin. Regularized estimation of Monge-Kantorovich quantiles for spherical data. *Journal of Computational and Graphical Statistics*, pages 1–31, 2026.

Ondrej Bojar, Christian Buck, Christian Federmann, et al. Findings of the 2014 Workshop on Statistical Machine Translation. In *WMT*, 2014.

Guillaume Carlier, Victor Chernozhukov, and Alfred Galichon. Vector quantile regression: an optimal transport approach. *The Annals of Statistics*, 44(3):1165–1192, 2016.

Victor Chernozhukov, Alfred Galichon, Marc Hallin, and Marc Henry. Monge-kantorovich depth, quantiles, ranks, and signs. *Annals of Statistics*, 2017.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013.

Jinhao Duan, Hao Cheng, Shiqi Wang, et al. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *ACL*, 2024.

Ivar Ekeland, Alfred Galichon, and Marc Henry. Comonotonic measures of multivariate risks. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics*, 22(1):109–132, 2012.

Falcon-LLM Team. The Falcon 3 family of open models, December 2024.

Yanqin Fan, Marc Henry, Brendan Pass, and Jorge A Rivero. Multidimensional inequality measurement via optimal transport. *Review of Economics and Statistics*, pages 1–45, 2024.

Olivier Paul Faugeras and Ludger Rüschendorf. Markov morphisms: a combined copula and mass transportation approach to multivariate quantiles. *Mathematica Applicanda*, 45(1), 2017.

Alexander Fishkov, Kajetan Schweighofer, Mykyta Ielanskyi, Nikita Kotelevskii, Mohsen Guizani, and Maxim Panov. Uncertainty quantification for regression using proper scoring rules. *Transactions on Machine Learning Research*, 2026.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, et al. Unsupervised Quality Estimation for Neural Machine Translation. *TACL*, 2020.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

Promit Ghosal and Bodhisattva Sen. Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing. *Annals of Statistics*, 2022.

Marc Hallin and Dimitri Konen. Multivariate quantiles: Geometric and measure-transportation-based contours. In Vladik Kreinovich, Woraphon Yamaka, and Supanika Leurcharusmee, editors, *Applications of Optimal Transport to Economics and Related Topics*, volume 556 of *Studies in Systems, Decision and Control*, pages 61–78. Springer, Cham, 2024.

Marc Hallin, Eustasio del Barrio, Juan Cuesta-Albertos, and Carlos Matrán. Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *Annals of Statistics*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.

Dan Hendrycks, Steven Basart, Norman Mu, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021.

Dan Hendrycks, Collin Burns, Steven Basart, et al. Measuring massive multitask language understanding. In *ICLR*, 2021.

Dan Hendrycks, Kevin Zhao, Steven Basart, et al. Natural adversarial examples. *CVPR*, 2021.

Paul Hofman, Yusuf Sale, and Eyke Hüllermeier. Uncertainty quantification for machine learning: One size does not fit all. In *AAAI*, 2026.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Chin-Wei Huang, Ricky TQ Chen, Christos Tsirigotis, and Aaron Courville. Convex potential flows: Universal probability distributions with optimal transport and convex optimization. In *ICLR*, 2021.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 2021.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*, 2017.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, 2017.

Michal Klein, Louis Béthune, Eugene Ndiaye, and Marco Cuturi. Multivariate conformal prediction using optimal transport. *Transactions on Machine Learning Research*, 2026.

Nikita Kotelevskii, Aleksandr Artemenkov, Kirill Fedyanin, et al. Nonparametric uncertainty quantification for single deterministic neural network. In *NeurIPS*, 2022.

Nikita Kotelevskii, Vladimir Kondratyev, Martin Takáč, et al. From risk to uncertainty: Generating predictive uncertainty measures via bayesian estimation. In *ICLR*, 2025.

A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Toronto*, 2009.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *ICLR*, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *TMLR*, 2023.

Andrey Malinin and Mark J. F. Gales. Uncertainty estimation in autoregressive structured prediction. In *ICLR*, 2021.

Meta AI. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Mistral AI Team. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *EMNLP*, 2018.

Yuval Netzer, Tao Wang, Adam Coates, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 2019.

Aram-Alexandre Pooladian and Jonathan Niles-Weed. Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*, 2021.

Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A conversational question answering challenge. *TACL*, 2019.

Ricardo Rei, José G. C. de Souza, Duarte Alves, et al. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *WMT*, 2022.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser, Cham, 2015.

Kajetan Schweighofer, Lukas Aichberger, Mykyta Ielanskyi, and Sepp Hochreiter. On information-theoretic measures of predictive uncertainty. *UAI*, 2025.

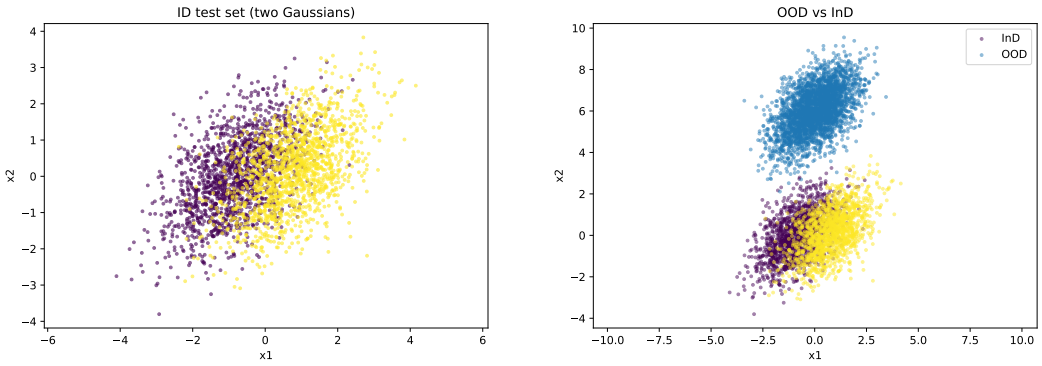
Gauthier Thurin, Kimia Nadjahi, and Claire Boyer. Optimal transport-based conformal prediction. In *ICML*, 2025.

Roman Vashurin, Maiya Goloburda, Albina Ilina, et al. Uncertainty quantification for llms through minimum bayes risk: Bridging confidence and consistency. In *NeurIPS*, 2025.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Lisa Wimmer, Yusuf Sale, Paul Hofman, et al. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *UAI*, 2023.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. Alignscore: Evaluating factual consistency with a unified alignment function. In *ACL*, 2023.



(a) In-distribution samples.

(b) In-distribution samples with OoD points.

Figure 5: Synthetic data used in the synthetic experiment.

A Motivating Synthetic Experiment

Here, we provide details of a synthetic experiment that illustrates robustness when only one component works well for a given downstream task.

We consider the following binary classification problem. Two Gaussians are centered at $(-0.8, 0.0)$ and $(0.8, 0.2)$, sharing covariance $\begin{bmatrix} 1.0 & 0.6 \\ 0.6 & 1.2 \end{bmatrix}$. Samples from the first Gaussian are labeled 0, and samples from the second are labeled 1. Data are shown in Figure 5 (left). As a model, we use logistic regression.

We evaluate two downstream UQ problems: *misclassification detection* and *out-of-distribution (OoD) detection*.

We compare two uncertainty measures: the maximum softmax probability (MSP) (Hendrycks and Gimpel, 2017) (recoverable as the zero-one proper scoring-rule instantiation of the Bayes risk (Kotelevskii *et al.*, 2025)) and the Mahalanobis score (Lee *et al.*, 2018).

The OoD data (another Gaussian) are placed so that the logistic regression is overly confident in assigning a class (Figure 5, right).

Accordingly, MSP performs well at misclassification detection but fails at OoD detection, whereas Mahalanobis performs the opposite. This is exactly what we observe in Figure 1 (main text). Our proposed composite, VecUQ-OT, is robust across both tasks. Numerical results appear in Table 6.

Scalar measures are typically task-specific. By contrast, combining complementary signals with VecUQ-OT yields robustness for both misclassification and OoD detection (see Figure 1).

Task	1-MSP	Mahalanobis	VecUQ-OT
Misclassification detection	0.766	0.445	0.710
OOD detection	0.099	1.000	1.000

Table 6: ROC-AUC on two downstream tasks: misclassification and OoD detection. Each measure (1-MSP or Mahalanobis) excels at only one task. Our vector-based combination (VecUQ-OT) remains robust across both.

We also illustrate, how the VecUQ-OT orders these 2-dimensional uncertainty vectors and assigns them to the Monge-Kantorovich ranks. In Figure 6, we see the distribution of the raw uncertainty scores (left), and the mapping to the uniform grid (right). The color of each point displays the VecUQ-OT rank assigned to this point. We see, that the resulting coloring is intuitive in a sense, that points with either of both high components have higher rank. Table 6 also reassures that the ordering is useful and robust in the considered downstream problem.

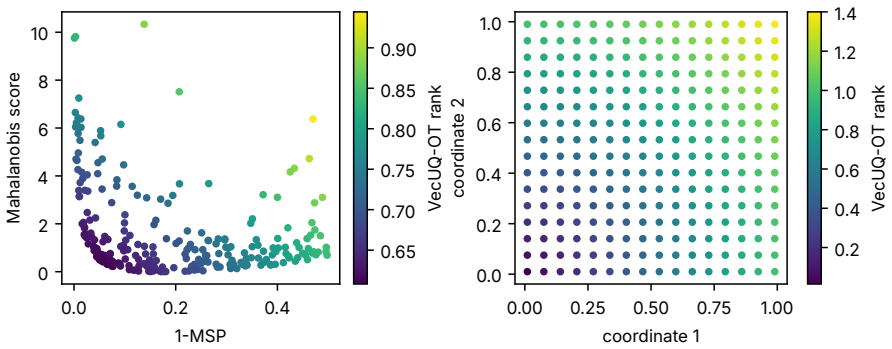


Figure 6: Ordering induced by VecUQ-OT in the experiment of Table 6. Left: calibration vectors (1-MSP, Mahalanobis), colored by their VecUQ-OT rank $\|\widehat{R}_\epsilon(x)\|_2$. Right: uniform reference points, colored by their norm, which is the rank of a vector mapped exactly onto them.

B Implementation

B.1 Limitations of Barycentric Projection and Implications for Our Method

By construction (see equation (2)), the rank image $\widehat{R}_\epsilon(x)$ (the barycentric projection) is a convex combination of the *target* atoms $\{\tilde{\mathbf{s}}_j\}$, i.e.,

$$\widehat{R}_\epsilon(x) \in \operatorname{conv}\{\tilde{\mathbf{s}}_j\}_{j=1}^n.$$

Consequently, once the coupling P^ϵ has been fit, the barycentric map cannot produce outputs outside the convex hull of the target support.

In our setting, the coupling is learned on *in-distribution* uncertainty vectors but must be applied to arbitrary inputs, including OoD ones. This means that mapped rank vectors for any new x satisfy $\widehat{R}_\epsilon(x) \in \operatorname{conv}\{\tilde{\mathbf{s}}_j\}_{\text{cal}}$. This leads to two effects:

1. *No extrapolation.* Inputs outside the calibration range cannot be mapped beyond the target hull; they are compressed toward it.
2. *Saturation far from the support.* Let $\mathbf{s} = c + tu$, where c is a fixed point (below, the mean of the calibration vectors), u is a unit direction, and $t > 0$. Expanding the square in the weights gives $\|\mathbf{s} - \tilde{\mathbf{s}}_j\|_2^2 = \|\mathbf{s}\|_2^2 - 2\langle \mathbf{s}, \tilde{\mathbf{s}}_j\rangle + \|\tilde{\mathbf{s}}_j\|_2^2$. The first term is the same for all j and cancels after normalization, so

$$w_j^\epsilon(\mathbf{s}) \propto v_j \exp\{(2\langle c, \tilde{\mathbf{s}}_j\rangle - \|\tilde{\mathbf{s}}_j\|_2^2)/\epsilon\} \exp\{2t\langle u, \tilde{\mathbf{s}}_j\rangle/\epsilon\}.$$

   Only the second factor depends on t. For two target points with $\langle u, \tilde{\mathbf{s}}_j\rangle > \langle u, \tilde{\mathbf{s}}_\ell\rangle$, the ratio $w_j^\epsilon(\mathbf{s})/w_\ell^\epsilon(\mathbf{s})$ is proportional to $\exp\{2t(\langle u, \tilde{\mathbf{s}}_j\rangle - \langle u, \tilde{\mathbf{s}}_\ell\rangle)/\epsilon\}$ and tends to infinity as $t \to \infty$. Hence, the weights concentrate on the target points with the largest projection $\langle u, \tilde{\mathbf{s}}_j\rangle$, and $\widehat{R}_\epsilon(x)$ approaches the boundary of the hull in the direction u. Its norm, which is the rank, converges to a finite value instead of growing with t.

Rank-based metrics are not affected as long as far OoD inputs are still ranked above in-distribution inputs. To check this, we take the mean of the calibration vectors $c = \frac{1}{N}\sum_{i=1}^N \mathbf{s}_i$ and push each OoD vector $\mathbf{s}$ away from it along the line $\mathbf{s}(t) = c + t(\mathbf{s} - c)$, where $t = 1$ gives the original vector. We then evaluate VecUQ-OT without refitting the OT map. Table 7 shows that ROC-AUC does not decrease.

B.2 Outer Anchors

In our experiments, we also add *outer anchor* points to the calibration vectors, placed beyond the calibration range, so that inputs outside this range are matched to target points nearer the boundary of $\operatorname{conv}\{\tilde{\mathbf{s}}_j\}$. For each coordinate $k \in \{1, \ldots, m\}$, let $M_k = \max_i s_{i,k}$ and fix $\gamma > 1$ (we use $\gamma = 5$). Then we add

$$\mathcal{A} = \Big\{a \in \mathbb{R}_+^m \colon\ a_k \in \{0, \gamma M_k\} \text{ for all } k,\ a \neq 0\Big\},$$

ROC-AUC	$t = 1$	$t = 1000$
CIFAR10 vs. SVHN	0.9593 ± 0.0033	0.9806 ± 0.0016
CIFAR100 vs. TinyImageNet	0.9999 ± 0.0000	1.0000 ± 0.0000
TinyImageNet vs. ImageNet-O	0.7522 ± 0.0026	0.8745 ± 0.0025

Table 7: OoD detection ROC-AUC when the OoD vectors are pushed away from the calibration data by a factor t, without refitting the OT map.

the $2^m - 1$ nonzero corners of the box $[0, \gamma M_1] \times \cdots \times [0, \gamma M_m]$. The anchors barely affect the results. Varying $\gamma \in \{1.5, 2, 5, 10, 20, 50\}$ changes the average ROC-AUC over the pairs of Table 1 by less than 10^{-5}. Using one anchor per axis instead of all $2^m - 1$ corners changes the mean ROC-AUC by less than 10^{-4} for m up to 6 and removes the exponential growth of the number of source points with m. Using no anchors matches the full set of corners within one standard deviation on every pair (see also Tables 8–11).

B.3 Reference Distribution and Discretization

We use two simple, isotropic factorized choices for the reference μ: (i) a product of exponentials (unbounded support on $\mathbb{R}_+^m$), and (ii) the uniform distribution on $[0, 1]^m$. Because we solve a *discrete* OT problem, we need target samples $\{\tilde{\mathbf{s}}_j\} \sim \mu$. To obtain them, we first draw points in $[0, 1]^m$ (a regular grid for $m \leq 2$ and Latin hypercube sampling for $m \geq 3$) and use it directly for the uniform reference, or transform each coordinate using the inverse CDF of the exponential distribution. Any strictly increasing radial reparameterization yields the same order, so we do not enforce normalization of rank vectors to the unit ball. Isotropy treats all coordinates symmetrically. Implementation details are in Appendix D.

B.4 Component Scaling

To improve the numerical stability of fitting entropy-regularized OT, we scale the components of $\mathbf{S}(x)$. We consider three options:

- **Global min-max scaling** (one scale for all components): preserves cross-component correlations.
- **Feature-wise min-max scaling** (per-component scales): may distort correlations but is more robust when component ranges differ greatly.
- **Feature-wise eCDF scaling**: replaces each component by its empirical CDF on the calibration set (Faugeras and Rüschendorf, 2017).

Our experiments showed that feature-wise min-max scaling performs better, especially when measures have disparate ranges. It is the default in all experiments. An ablation study of design choices and a discussion of limitations are provided in Section C and Section E, respectively.

With all these design choices, we refer to the procedure as **VecUQ-OT**.

C Ablation on Design Choices

We study how three design factors affect performance: (i) the *target distribution* for the OT reference cloud (Uniform on $[0, 1]^m$ vs. Exponential on $\mathbb{R}_+^m$), (ii) the *component scaling* (Feature-wise = per-feature min–max, Global = one min–max over all features, Identity = no scaling), and (iii) the *outer-anchor scale* $\gamma \in \{5, 2, 0\}$ *of Appendix B.2* (where 0 means no anchors).

We evaluate OoD detection on two ID/OoD pairs: CIFAR10 vs. CIFAR100 (Krizhevsky, 2009) and TinyImageNet (Le and Yang, 2015) vs. ImageNet-R (Hendrycks *et al.*, 2021a). As a composite uncertainty vector, we use the same type of combination as in Section 5.2 (Bayes, excess, and total risks with the Mahalanobis score), here with the approximations $R_b^1, R_{\text{exc}}^{1,1}, R_{\text{tot}}^{1,1}, \text{MahS}$. Results are in Table 8; we use four ensemble groups of five members and report mean $\pm$ std ROC-AUC. From Table 8, results are close overall, with Feature-wise scaling typically strongest.

InD	OOD	ROC AUC	Target	ScalingType	γ
CIFAR10	CIFAR100	0.917984 ± 0.000778	Uniform	FeatureWise	5
CIFAR10	CIFAR100	0.917983 ± 0.000779	Uniform	FeatureWise	2
CIFAR10	CIFAR100	0.917950 ± 0.000783	Uniform	FeatureWise	0
CIFAR10	CIFAR100	0.917474 ± 0.000840	Exp	FeatureWise	0
CIFAR10	CIFAR100	0.917423 ± 0.000835	Exp	FeatureWise	2
CIFAR10	CIFAR100	0.917394 ± 0.000832	Exp	FeatureWise	5
CIFAR10	CIFAR100	0.915669 ± 0.000847	Exp	Identity	5
CIFAR10	CIFAR100	0.915552 ± 0.000814	Exp	Identity	2
CIFAR10	CIFAR100	0.915245 ± 0.001195	Exp	Identity	0
CIFAR10	CIFAR100	0.913731 ± 0.001273	Uniform	Global	5
CIFAR10	CIFAR100	0.913727 ± 0.001273	Uniform	Global	2
CIFAR10	CIFAR100	0.913700 ± 0.001274	Uniform	Global	0
CIFAR10	CIFAR100	0.913466 ± 0.001299	Exp	Global	5
CIFAR10	CIFAR100	0.913454 ± 0.001300	Exp	Global	2
CIFAR10	CIFAR100	0.913383 ± 0.001302	Exp	Global	0
CIFAR10	CIFAR100	0.910275 ± 0.000686	Uniform	Identity	0
CIFAR10	CIFAR100	0.907669 ± 0.000635	Uniform	Identity	2
CIFAR10	CIFAR100	0.907524 ± 0.000649	Uniform	Identity	5

Table 8: Ablation on target distribution, scaling, and outer-anchor scale γ for CIFAR10 vs. CIFAR100. Composite: $R_b^1, R_{\text{exc}}^{1,1}, R_{\text{tot}}^{1,1}, \text{MahS}$. Mean $\pm$ std ROC–AUC over four ensembles of five members.

Next, we consider a different composition: various excess-risk approximations with log-score instantiation, $R_{\text{exc}}^{1,1}, R_{\text{exc}}^{1,2}, R_{\text{exc}}^{2,1}, R_{\text{exc}}^{1,3}, R_{\text{exc}}^{3,1}$. Results (Table 9) are again close overall; the Uniform target with Feature-wise scaling performs best.

InD	OOD	ROC AUC	Target	ScalingType	γ
CIFAR10	CIFAR100	0.904907 ± 0.000304	Uniform	FeatureWise	0
CIFAR10	CIFAR100	0.904833 ± 0.000305	Uniform	FeatureWise	5
CIFAR10	CIFAR100	0.904826 ± 0.000306	Uniform	FeatureWise	2
CIFAR10	CIFAR100	0.904776 ± 0.000290	Uniform	Identity	0
CIFAR10	CIFAR100	0.904726 ± 0.000288	Uniform	Global	0
CIFAR10	CIFAR100	0.904714 ± 0.000295	Uniform	Identity	5
CIFAR10	CIFAR100	0.904710 ± 0.000295	Uniform	Identity	2
CIFAR10	CIFAR100	0.904683 ± 0.000292	Uniform	Global	5
CIFAR10	CIFAR100	0.904681 ± 0.000292	Uniform	Global	2
CIFAR10	CIFAR100	0.904659 ± 0.000298	Exp	FeatureWise	0
CIFAR10	CIFAR100	0.904621 ± 0.000313	Exp	FeatureWise	2
CIFAR10	CIFAR100	0.904621 ± 0.000313	Exp	FeatureWise	5
CIFAR10	CIFAR100	0.904593 ± 0.000289	Exp	Global	0
CIFAR10	CIFAR100	0.904569 ± 0.000296	Exp	Global	2
CIFAR10	CIFAR100	0.904567 ± 0.000296	Exp	Global	5
CIFAR10	CIFAR100	0.904473 ± 0.000278	Exp	Identity	0
CIFAR10	CIFAR100	0.904417 ± 0.000289	Exp	Identity	5
CIFAR10	CIFAR100	0.904413 ± 0.000288	Exp	Identity	2

Table 9: Ablation with excess-risk variants for CIFAR10 vs. CIFAR100. Composite: $R_{\text{exc}}^{1,1}, R_{\text{exc}}^{1,2}, R_{\text{exc}}^{2,1}, R_{\text{exc}}^{1,3}, R_{\text{exc}}^{3,1}$. Mean $\pm$ std ROC–AUC over four ensembles of five members.

For TinyImageNet vs. ImageNet-R, we evaluate the same combinations. Results are in Tables 10 and 11. In Table 10, Identity scaling is omitted because the entropic-OT fit produced NaNs due to components with highly disparate magnitudes (e.g., MahS). We also observe that Global scaling performs poorly for this dataset/measure choice, whereas the Uniform target with Feature-wise scaling remains robust. Overall, the observations are consistent with the previous ID/OOD pair.

Scaling by the empirical CDF. Table 12 compares feature-wise min-max scaling with feature-wise eCDF scaling, both fitted on the same unlabeled calibration data and applied before OT, on the composition of Table 1. The eCDF scaling is competitive, but slightly worse on average. The two scalings differ in how they treat values outside the calibration range. Min-max scaling is linear and preserves relative distances between observations, so new values may exceed 1. The eCDF keeps only the order of the values and maps all values above the calibration maximum to 1.

InD	OOD	ROC AUC	Target	ScalingType	γ
TinyImageNet	ImageNet-R	0.834519 ± 0.000639	Uniform	FeatureWise	5
TinyImageNet	ImageNet-R	0.834514 ± 0.000639	Uniform	FeatureWise	2
TinyImageNet	ImageNet-R	0.834509 ± 0.000652	Uniform	FeatureWise	0
TinyImageNet	ImageNet-R	0.831367 ± 0.000777	Exp	FeatureWise	0
TinyImageNet	ImageNet-R	0.831306 ± 0.000835	Exp	FeatureWise	2
TinyImageNet	ImageNet-R	0.831289 ± 0.000829	Exp	FeatureWise	5
TinyImageNet	ImageNet-R	0.421999 ± 0.006611	Exp	Global	5
TinyImageNet	ImageNet-R	0.421810 ± 0.006610	Exp	Global	2
TinyImageNet	ImageNet-R	0.420473 ± 0.006612	Exp	Global	0
TinyImageNet	ImageNet-R	0.419511 ± 0.006852	Uniform	Global	5
TinyImageNet	ImageNet-R	0.419462 ± 0.006853	Uniform	Global	2
TinyImageNet	ImageNet-R	0.419213 ± 0.006850	Uniform	Global	0

Table 10: Ablation on target distribution, scaling, and outer-anchor scale γ for TinyImageNet vs. ImageNet-R. Composite: $R_b^1, R_{\text{exc}}^{1,1}, R_{\text{tot}}^{1,1}, \text{MahS}$. Mean ± std ROC–AUC over four ensembles of five members.

InD	OOD	ROC AUC	Target	ScalingType	γ
TinyImageNet	ImageNet-R	0.777316 ± 0.003418	Uniform	FeatureWise	5
TinyImageNet	ImageNet-R	0.777304 ± 0.003420	Uniform	FeatureWise	2
TinyImageNet	ImageNet-R	0.777266 ± 0.003377	Uniform	FeatureWise	0
TinyImageNet	ImageNet-R	0.776733 ± 0.003649	Exp	FeatureWise	2
TinyImageNet	ImageNet-R	0.776686 ± 0.003647	Exp	FeatureWise	5
TinyImageNet	ImageNet-R	0.776372 ± 0.003673	Exp	FeatureWise	0
TinyImageNet	ImageNet-R	0.774565 ± 0.002815	Uniform	Identity	0
TinyImageNet	ImageNet-R	0.774534 ± 0.002825	Uniform	Identity	5
TinyImageNet	ImageNet-R	0.774532 ± 0.002832	Uniform	Identity	2
TinyImageNet	ImageNet-R	0.774266 ± 0.002822	Uniform	Global	5
TinyImageNet	ImageNet-R	0.774248 ± 0.002823	Uniform	Global	2
TinyImageNet	ImageNet-R	0.774174 ± 0.002803	Uniform	Global	0
TinyImageNet	ImageNet-R	0.774099 ± 0.002938	Exp	Global	2
TinyImageNet	ImageNet-R	0.774089 ± 0.002924	Exp	Global	5
TinyImageNet	ImageNet-R	0.773737 ± 0.002870	Exp	Global	0
TinyImageNet	ImageNet-R	0.773190 ± 0.002905	Exp	Identity	2
TinyImageNet	ImageNet-R	0.773148 ± 0.002895	Exp	Identity	5
TinyImageNet	ImageNet-R	0.772700 ± 0.002874	Exp	Identity	0

Table 11: Ablation with excess-risk variants for TinyImageNet vs. ImageNet-R. Composite: $R_{\text{exc}}^{1,1}, R_{\text{exc}}^{1,2}, R_{\text{exc}}^{2,1}, R_{\text{exc}}^{1,3}, R_{\text{exc}}^{3,1}$. Mean ± std ROC–AUC over four ensembles of five members.

D Details on Sampling from Reference Distribution

Our OT-based ordering maps calibration vectors $\mathbf{s}_i \in \mathbb{R}_+^m$ to a simple, factorized reference distribution μ and ranks points by radius in the reference space (via barycentric projection). Because we solve a *discrete* entropic OT problem, we approximate μ with a finite cloud of target points $\{\tilde{\mathbf{s}}_j\}_{j=1}^n$. This section details how we sample those targets.

We first draw points $U \in [0,1]^{n\times m}$ in the unit hypercube, using a regular grid for $m \leq 2$ and, for $m \geq 3$, a stratified random sample in which every coordinate takes exactly one value in each of n equal intervals of $[0,1]$. Then, we transform each coordinate of U with the inverse CDF of the chosen marginal, yielding independent coordinates and a factorized μ:

- **Product of exponentials.** For rates $\lambda_\ell > 0$,

$$\tilde{\mathbf{s}}_{j\ell} = F_{\text{Exp}(\lambda_\ell)}^{-1}(U_{j\ell}) = -\frac{1}{\lambda_\ell}\log\big(1 - U_{j\ell}\big),$$

  supported on $\mathbb{R}_+^m$. We use a common rate $\lambda_\ell \equiv \lambda$ for all coordinates.
- **Uniform distribution on $[0,1]^m$.** Here the inverse CDF is the identity, so $\tilde{\mathbf{s}}_{j\ell} = U_{j\ell}$.

E Limitations

We highlight several limitations of our approach.

1. **Need for a separate calibration set.** Although the OT-based mapping is lightweight to fit, it requires a separate *covariates-only* dataset for calibration. Ideally, this set would include both in-distribution and OoD samples, allowing us to observe uncertainty vectors

InD vs. OoD	Min-max (default)	eCDF
CIFAR10 vs. CIFAR100	0.919 ± 0.001	0.920 ± 0.001
CIFAR10 vs. SVHN	0.959 ± 0.003	0.958 ± 0.003
CIFAR10 vs. TinyImageNet	0.914 ± 0.001	0.915 ± 0.001
CIFAR100 vs. CIFAR10	0.767 ± 0.001	0.756 ± 0.001
CIFAR100 vs. SVHN	0.873 ± 0.005	0.875 ± 0.010
CIFAR100 vs. TinyImageNet	1.000 ± 0.000	0.928 ± 0.002
TinyImageNet vs. ImageNet-A	0.847 ± 0.002	0.828 ± 0.004
TinyImageNet vs. ImageNet-R	0.834 ± 0.002	0.805 ± 0.003
TinyImageNet vs. ImageNet-O	0.752 ± 0.003	0.758 ± 0.002
Average	0.874 ± 0.002	0.860 ± 0.003

Table 12: OoD ROC-AUC of VecUQ-OT with feature-wise min-max and eCDF scaling (mean $\pm$ std over ensemble groups).

In-Dist.	Evaluation	100%	70%	50%	20%	10%	1%	0.5%
CIFAR-10	CIFAR-10 [miscls]	.944 ± .002	.944 ± .002	.944 ± .002	.944 ± .002	.944 ± .002	.944 ± .002	.941 ± .002
	CIFAR-10 [selective]	.997 ± .000	.997 ± .000	.997 ± .000	.997 ± .000	.997 ± .000	.997 ± .000	.997 ± .000
	CIFAR-100 [ood]	.918 ± .001	.918 ± .001	.918 ± .001	.918 ± .001	.918 ± .001	.917 ± .002	.916 ± .003
	SVHN [ood]	.957 ± .005	.957 ± .005	.957 ± .005	.958 ± .005	.958 ± .005	.956 ± .006	.952 ± .006
	Tiny-ImageNet [ood]	.912 ± .001	.912 ± .000	.912 ± .000	.913 ± .001	.912 ± .001	.911 ± .002	.910 ± .004
CIFAR-100	CIFAR-10 [ood]	.765 ± .001	.764 ± .001	.763 ± .001	.764 ± .002	.760 ± .003	.762 ± .003	.634 ± .021
	CIFAR-100 [miscls]	.849 ± .003	.849 ± .002	.848 ± .003	.848 ± .004	.846 ± .004	.848 ± .002	.703 ± .031
	CIFAR-100 [selective]	.916 ± .001	.915 ± .001	.915 ± .001	.915 ± .001	.914 ± .001	.916 ± .001	.867 ± .010
	SVHN [ood]	.870 ± .006	.870 ± .006	.868 ± .007	.870 ± .007	.864 ± .006	.854 ± .006	.738 ± .018
	Tiny-ImageNet [ood]	1.00 ± .000	1.00 ± .000	1.00 ± .000	1.00 ± .000	1.00 ± .000	1.00 ± .000	.995 ± .006
Tiny-ImageNet	Tiny-ImageNet [miscls]	.847 ± .002	.843 ± .002	.841 ± .003	.843 ± .003	.839 ± .003	.842 ± .004	.720 ± .018
	Tiny-ImageNet [selective]	.886 ± .001	.884 ± .001	.883 ± .001	.884 ± .001	.882 ± .002	.885 ± .001	.835 ± .007
	ImageNet-A [ood]	.847 ± .002	.846 ± .002	.845 ± .002	.846 ± .002	.843 ± .002	.840 ± .005	.744 ± .014
	ImageNet-R [ood]	.835 ± .001	.832 ± .001	.831 ± .001	.832 ± .001	.830 ± .001	.828 ± .006	.727 ± .016
	ImageNet-O [ood]	.760 ± .002	.762 ± .001	.763 ± .002	.762 ± .002	.763 ± .002	.764 ± .003	.712 ± .010

Table 13: VecUQ-OT with the OT map fitted on a fraction of the calibration set. 100% is the full calibration set of the main text (1800 inputs). For OoD detection [ood] and misclassification detection [miscls], we report ROC-AUC, and for selective prediction [selective], the area under the accuracy-coverage curve.

where OoD-typical scores lie. However, relying on explicit OoD data reduces generality. Consequently, in our experiments, we assume no OoD data and calibrate on in-distribution samples only. The outer anchors of Appendix B.2 have almost no effect on the results.

2. **Choice of cost function.** We use squared Euclidean distance as the OT cost. While empirically effective, it may not be optimal for transferring one uncertainty representation to another. Exploring alternatives within entropy-regularized, discrete OT is nontrivial and may complicate convergence and stability.

3. **Radial ranking discards direction.** Our final scalar score depends only on the norm of the barycentric image (a radial rank), not its direction. Although the isotropic reference is designed to make angles uninformative, directional information can still carry task-specific signals that our current rank ignores.

F Ablation on Calibration Dataset Size

In this section, we study how the size of the calibration dataset used for fitting the OT map affects the results.

To study this, we varied the size of the calibration data (100% corresponds to the results presented in the main text of the paper, with 18% (=1800) objects across all datasets). See the Table 13 below for the results.

We observe that performance remains stable under substantial reductions of the calibration set size, with degradation mainly appearing only in the most extreme 0.5% setting.

G Comparison with Additive Aggregation

VecUQ-OT allows nonlinear aggregation of different uncertainty measures, which is not equivalent to simply summing them. To demonstrate this, we conducted an experiment comparing the additive TU measure (which aggregates AU and EU estimates by summing them) with the same components aggregated with VecUQ-OT. The results are presented in Table 14.

From the results, we see that the ordering indeed changes across all datasets. While the performance for these two combinations is close, our method provides a slight improvement on some datasets. Below, we explain why the order changed.

Since our approach allows for non-additive aggregation of AU and EU, it will not be equivalent to simply adding them. From the construction of the final uncertainty measure, the resulting ordering will be determined as a radial distance from the origin to the image of the transport map, which is a nonlinear aggregation.

The additive aggregation of uncertainty, although theoretically correct, has recently been criticized (Wimmer *et al.*, 2023) (see Section 4.4 therein) for failing to hold in practical machine learning scenarios. Our approach is one of the possible methods for achieving non-additive aggregation.

The estimates of AU and EU can be obtained in arbitrary ways, and in principle, they can even be of different natures (AU from an information-theoretic perspective, EU from a density-based perspective). In this case, when AU and EU estimates have different natures and scales, additive aggregation requires tuning the aggregation coefficients. In our case, everything is fitted in an unsupervised manner.

In-Dist.	**OOD Dataset**	**Method**	**Score (↑)**
CIFAR-10	CIFAR-100	VecUQ-OT	**.9138** $\pm$ .0006
		TU (sum)	.9115 $\pm$ .0005
Tiny-ImageNet	ImageNet-O	VecUQ-OT	**.7565** $\pm$ .0031
		TU (sum)	.7535 $\pm$ .0017
	ImageNet-R	VecUQ-OT	.8360 $\pm$ .0011
		TU (sum)	**.8366** $\pm$ .0015

Table 14: OoD detection ROC-AUC for the same AU and EU components, aggregated either by VecUQ-OT or by summation (TU). Best in bold.

H Extreme Compositions

We examine edge cases to test the robustness and sanity of our vector-based aggregation. In each case, VecUQ-OT should behave predictably: it must *not* invent structure when none is present and should preserve useful ordering when a signal exists. Results in this section are consistent across the OT design choices discussed in the main text (reference distribution, component normalization, etc.).

H.1 Single-component Vector

We set the composite vector to contain a *single* uncertainty measure as a sanity check. In this setting, VecUQ-OT should reproduce the original ordering exactly, since there is no multivariate structure to exploit. For example, Table 15 reports OOD ROC-AUC on CIFAR10 (ID) vs. CIFAR100 (OOD) using an excess-risk estimator (log-score instantiation (Kotelevskii *et al.*, 2025)). The ordering and ROC-AUC are unchanged after applying VecUQ-OT.

In-distribution	Out-of-distribution	Measure	Mean ROC AUC	Std ROC AUC
CIFAR10	CIFAR100	$R^{1,1}_{\text{exc}}$	0.9047	0.0003
CIFAR10	CIFAR100	VecUQ-OT	0.9047	0.0003

Table 15: Single-component composition: the vector contains only $R^{1,1}_{\text{exc}}$. VecUQ-OT preserves the original ordering (identical ROC-AUC). Ensembles: 4 groups of 5 members.

H.2 Stacking Identical Measures

Next, we stack the *same* measure multiple times (duplicate coordinates). The vector carries no additional information beyond the single measure, so the induced order should match the base measure again. Table 16 confirms this.

In-distribution	Out-of-distribution	Measure	Mean ROC AUC	Std ROC AUC
CIFAR10	CIFAR100	$R_{\text{exc}}^{1,1}$	0.9047	0.0003
CIFAR10	CIFAR100	$R_{\text{exc}}^{1,1}$	0.9047	0.0003
CIFAR10	CIFAR100	$R_{\text{exc}}^{1,1}$	0.9047	0.0003
CIFAR10	CIFAR100	$R_{\text{exc}}^{1,1}$	0.9047	0.0003
CIFAR10	CIFAR100	$R_{\text{exc}}^{1,1}$	0.9047	0.0003
CIFAR10	CIFAR100	VecUQ-OT	0.9047	0.0003

Table 16: Duplicated-measure composition: the vector repeats $R_{\text{exc}}^{1,1}$ across coordinates. VecUQ-OT leaves the ranking unchanged, as expected. Ensembles: 4 groups of 5 members.

H.3 One Signal, the Rest Constants

Finally, we build a vector where *one* coordinate carries the signal and all others are constants. A sensible aggregator should ignore the constant coordinates and follow the informative ones. Table 17 shows that VecUQ-OT matches the ROC-AUC of the informative measure and is unaffected by constant dimensions.

In-distribution	Out-of-distribution	Metric	Mean ROC AUC	Std ROC AUC
CIFAR10	CIFAR100	$R_{\text{exc}}^{1,1}$	0.9047	0.0003
CIFAR10	CIFAR100	Constant 1	0.5000	0.0000
CIFAR10	CIFAR100	Constant 2	0.5000	0.0000
CIFAR10	CIFAR100	Constant 3	0.5000	0.0000
CIFAR10	CIFAR100	Constant 4	0.5000	0.0000
CIFAR10	CIFAR100	Constant 5	0.5000	0.0000
CIFAR10	CIFAR100	VecUQ-OT	0.9047	0.0003

Table 17: One-signal composition: the vector contains $R_{\text{exc}}^{1,1}$ plus several constant coordinates. VecUQ-OT follows the informative coordinate and ignores constants. Ensembles: 4 groups of 5 members.

I Other Measure Choices

In this section, we examine additional combinations of uncertainty measures.

Aggregation of different total-uncertainty estimates. *Total* uncertainty is typically decomposed into aleatoric and epistemic parts, aims to capture all sources of predictive uncertainty, and is often a safe default for downstream tasks. However, as shown in Kotelevskii *et al.* (2025); Schweighofer *et al.* (2025); Hofman *et al.* (2026), there are multiple valid instantiations via proper scoring rules and several Bayesian approximation strategies for the total uncertainty estimate. Here we fix one Bayesian estimate, $R_{\text{tot}}^{1,2}$, and follow the setup from Section 5.2 with four deep-ensemble groups of five members each. We then instantiate the total-risk estimate with several proper scoring rules (Logscore, Brier, Spherical, and 0-1) and form a composite vector from these components. Table 18 shows that VecUQ-OT is robust: on average, it performs best across problems and, in most cases, ranks best or second-best. Among individual measures, the Logscore instantiation often performs best, but it can occasionally underperform (e.g., CIFAR100 and TinyImageNet for misclassification detection and selective prediction), whereas VecUQ-OT remains stable.

Aggregation of different aleatoric-uncertainty estimates. *Aleatoric* uncertainty is used to detect regions with increased label noise. As shown in Kotelevskii *et al.* (2025); Schweighofer *et al.* (2025), it is also effective for OOD detection. Table 19 reports results for aggregating different R_b^1 estimates. The pattern mirrors Table 18: the Logscore-based measure most frequently attains the top score (and here is best on average), but for some problems it drops to the bottom. In contrast, our composite measure is very robust, never the worst and typically best or second-best.

Aggregation of different epistemic-uncertainty estimates. *Epistemic* uncertainty targets inputs where the model lacks sufficient knowledge of the data-generating process, which includes OOD

InD	Eval	VecUQ-OT	Logscore	Brier	Spherical	Zero-one
CIFAR10	CIFAR10 [miscls]	**0.946 ± 0.002**	0.945 ± 0.002	**0.946 ± 0.002**	**0.946 ± 0.002**	**0.946 ± 0.002**
	CIFAR10 [selective]	**0.997 ± 0.000**	**0.997 ± 0.000**	**0.997 ± 0.000**	**0.997 ± 0.000**	**0.997 ± 0.000**
	CIFAR100 [ood]	0.913 ± 0.001	**0.916 ± 0.001**	0.913 ± 0.001	0.913 ± 0.001	0.911 ± 0.001
	SVHN [ood]	0.958 ± 0.003	**0.963 ± 0.003**	0.958 ± 0.004	0.958 ± 0.004	0.955 ± 0.004
	TinyImageNet [ood]	0.906 ± 0.001	**0.910 ± 0.001**	0.906 ± 0.001	0.906 ± 0.001	0.904 ± 0.001
CIFAR100	CIFAR10 [ood]	0.774 ± 0.002	**0.775 ± 0.002**	0.774 ± 0.002	0.774 ± 0.002	0.771 ± 0.001
	CIFAR100 [miscls]	0.864 ± 0.003	0.850 ± 0.003	0.866 ± 0.003	0.866 ± 0.003	**0.870 ± 0.003**
	CIFAR100 [selective]	0.922 ± 0.001	0.918 ± 0.001	0.922 ± 0.001	0.922 ± 0.001	**0.924 ± 0.001**
	SVHN [ood]	0.859 ± 0.006	**0.870 ± 0.006**	0.857 ± 0.006	0.857 ± 0.006	0.847 ± 0.006
	TinyImageNet [ood]	0.938 ± 0.001	0.926 ± 0.001	0.937 ± 0.001	0.937 ± 0.001	**0.947 ± 0.002**
TinyImageNet	TinyImageNet [miscls]	0.863 ± 0.003	0.850 ± 0.003	0.863 ± 0.003	0.863 ± 0.003	**0.868 ± 0.004**
	TinyImageNet [selective]	0.895 ± 0.001	0.891 ± 0.000	0.896 ± 0.001	0.896 ± 0.001	**0.897 ± 0.001**
	ImageNet-A [ood]	0.836 ± 0.003	**0.841 ± 0.002**	0.834 ± 0.003	0.834 ± 0.003	0.827 ± 0.003
	ImageNet-R [ood]	0.825 ± 0.003	**0.831 ± 0.003**	0.824 ± 0.003	0.824 ± 0.003	0.817 ± 0.003
	ImageNet-O [ood]	**0.736 ± 0.003**	0.735 ± 0.003	**0.736 ± 0.003**	**0.736 ± 0.003**	0.735 ± 0.004
Mean		0.8821 ± 0.002	0.8811 ± 0.002	0.8819 ± 0.002	0.8819 ± 0.002	0.8809 ± 0.002

Table 18: Performance for compositions of different approximations of $R_{\text{tot}}^{1,2}$. We consider all problems from Sec. 5.2. Here, [miscls] denotes misclassification detection, [selective] selective prediction, and [ood] OOD detection. For [miscls] and [ood] we report ROC–AUC; for [selective] we report area under the accuracy–coverage curve.

InD	Eval	VecUQ-OT	Logscore	Brier	Spherical	Zero-one
CIFAR10	CIFAR10 [miscls]	**0.942 ± 0.002**	**0.942 ± 0.002**	**0.942 ± 0.002**	**0.942 ± 0.002**	**0.942 ± 0.002**
	CIFAR10 [selective]	**0.997 ± 0.000**	**0.997 ± 0.000**	**0.997 ± 0.000**	**0.997 ± 0.000**	**0.997 ± 0.000**
	CIFAR100 [ood]	0.915 ± 0.001	**0.917 ± 0.001**	0.914 ± 0.001	0.915 ± 0.001	0.913 ± 0.001
	SVHN [ood]	0.959 ± 0.002	**0.963 ± 0.002**	0.958 ± 0.002	0.959 ± 0.002	0.956 ± 0.002
	TinyImageNet [ood]	0.909 ± 0.001	**0.911 ± 0.001**	0.909 ± 0.001	0.909 ± 0.001	0.907 ± 0.001
CIFAR100	CIFAR10 [ood]	**0.773 ± 0.002**	**0.773 ± 0.002**	**0.773 ± 0.002**	**0.773 ± 0.002**	0.772 ± 0.002
	CIFAR100 [miscls]	0.855 ± 0.003	0.845 ± 0.003	0.858 ± 0.003	0.856 ± 0.003	**0.859 ± 0.003**
	CIFAR100 [selective]	0.919 ± 0.001	0.916 ± 0.001	0.920 ± 0.001	0.920 ± 0.001	**0.921 ± 0.001**
	SVHN [ood]	0.862 ± 0.006	**0.870 ± 0.007**	0.858 ± 0.006	0.861 ± 0.006	0.856 ± 0.006
	TinyImageNet [ood]	0.803 ± 0.001	**0.810 ± 0.001**	0.790 ± 0.001	0.806 ± 0.001	0.803 ± 0.000
TinyImageNet	TinyImageNet [miscls]	0.853 ± 0.003	0.845 ± 0.003	**0.855 ± 0.003**	0.853 ± 0.003	**0.855 ± 0.003**
	TinyImageNet [selective]	0.892 ± 0.001	0.889 ± 0.001	**0.893 ± 0.001**	0.892 ± 0.001	**0.893 ± 0.001**
	ImageNet-A [ood]	0.831 ± 0.003	**0.835 ± 0.002**	0.827 ± 0.003	0.830 ± 0.003	0.826 ± 0.003
	ImageNet-R [ood]	0.820 ± 0.003	**0.825 ± 0.003**	0.816 ± 0.003	0.819 ± 0.003	0.815 ± 0.003
	ImageNet-O [ood]	0.723 ± 0.004	**0.724 ± 0.004**	0.721 ± 0.004	0.723 ± 0.004	0.721 ± 0.004
AVG	[all rows]	0.870 ± 0.002	0.871 ± 0.002	0.869 ± 0.002	0.870 ± 0.002	0.869 ± 0.002

Table 19: Performance for compositions of different approximations of R_b^1. We consider all problems from Sec. 5.2. Here, [miscls] denotes misclassification detection, [selective] selective prediction, and [ood] OOD detection. For [miscls] and [ood] we report ROC–AUC; for [selective] we report area under the accuracy–coverage curve.

detection. In this experiment, we aggregate several instantiations (induced by different proper scoring rules) of the excess-risk measure $R_{\text{exc}}^{1,3}$. Results are shown in Table 20. Consistent with previous observations, VecUQ-OT is highly robust. Logscore and Spherical perform well on average, but each is sometimes near the bottom, whereas VecUQ-OT is typically best or second-best.

J Textual Experiments Details

Datasets. We use datasets covering several NLP tasks in our experimental setup, including summarization, translation, and long- and short-form question answering. For summarization, we use the XSum dataset (Narayan *et al.*, 2018). We use the WMT14 Fr-En and WMT19 De-En datasets (Bojar *et al.*, 2014; Barrault *et al.*, 2019) for translation. For short-form question answering, we use the MMLU and Gsm8k datasets (Hendrycks *et al.*, 2021b; Cobbe *et al.*, 2021), while for long-form question answering, we employ TriviaQA and CoQA (Joshi *et al.*, 2017; Reddy *et al.*, 2019).

Models. For our experiments, we use the base versions of LLaMA-3.1 8B, Mistral-7B, and Falcon-7B (Meta AI, 2024; Mistral AI Team, 2023; Falcon-LLM Team, 2024).

Evaluation. We solve the task of selective generation by using the uncertainty score as a proxy for quality and rejecting outputs based on it. However, traditional metrics such as ROC-AUC or ECE are not applicable to textual outputs. Not only do some uncertainty-estimation methods produce unbounded values, but generation quality is also non-discrete. Thus, we use the Prediction Rejection Ratio (PRR) (Malinin and Gales, 2021). PRR assesses how the average quality of generated outputs

InD	Eval	VecUQ-OT	Logscore	Brier	Spherical	Zero-one
CIFAR10	CIFAR10 [miscls]	**0.942 ± 0.003**	0.936 ± 0.003	**0.942 ± 0.003**	**0.942 ± 0.003**	0.797 ± 0.008
	CIFAR10 [selective]	**0.997 ± 0.000**	**0.997 ± 0.000**	**0.997 ± 0.000**	**0.997 ± 0.000**	0.983 ± 0.002
	CIFAR100 [ood]	**0.904 ± 0.000**	0.903 ± 0.001	0.902 ± 0.000	**0.904 ± 0.000**	0.755 ± 0.001
	SVHN [ood]	0.941 ± 0.011	0.940 ± 0.013	0.940 ± 0.010	**0.942 ± 0.010**	0.825 ± 0.038
	TinyImageNet [ood]	**0.895 ± 0.000**	0.894 ± 0.001	0.893 ± 0.000	0.894 ± 0.001	0.752 ± 0.002
CIFAR100	CIFAR10 [ood]	0.710 ± 0.002	**0.718 ± 0.002**	0.681 ± 0.002	0.715 ± 0.002	0.689 ± 0.002
	CIFAR100 [miscls]	0.819 ± 0.003	0.804 ± 0.003	0.783 ± 0.005	**0.827 ± 0.003**	0.806 ± 0.003
	CIFAR100 [selective]	0.910 ± 0.001	0.906 ± 0.001	0.900 ± 0.002	**0.914 ± 0.001**	0.879 ± 0.003
	SVHN [ood]	0.718 ± 0.011	**0.750 ± 0.016**	0.662 ± 0.007	0.719 ± 0.010	0.706 ± 0.011
	TinyImageNet [ood]	0.995 ± 0.001	**1.000 ± 0.000**	0.953 ± 0.003	0.990 ± 0.001	0.976 ± 0.002
TinyImageNet	TinyImageNet [miscls]	0.799 ± 0.002	0.791 ± 0.002	0.754 ± 0.003	**0.805 ± 0.001**	0.801 ± 0.004
	TinyImageNet [selective]	0.875 ± 0.001	0.872 ± 0.001	0.860 ± 0.002	**0.878 ± 0.002**	0.853 ± 0.003
	ImageNet-A [ood]	0.716 ± 0.003	**0.772 ± 0.004**	0.651 ± 0.003	0.715 ± 0.002	0.713 ± 0.002
	ImageNet-R [ood]	0.715 ± 0.002	**0.764 ± 0.003**	0.657 ± 0.005	0.716 ± 0.004	0.712 ± 0.003
	ImageNet-O [ood]	0.721 ± 0.003	**0.746 ± 0.003**	0.691 ± 0.005	0.719 ± 0.003	0.700 ± 0.004
AVG	[all rows]	0.844 ± 0.003	0.853 ± 0.003	0.818 ± 0.003	0.845 ± 0.003	0.796 ± 0.006

Table 20: Performance for compositions of different approximations of $R^{1,3}_{\text{exc}}$. We consider all problems from Sec. 5.2. Here, [miscls] denotes misclassification detection, [selective] selective prediction, and [ood] OOD detection. For [miscls] and [ood] we report ROC–AUC; for [selective] we report area under the accuracy–coverage curve.

changes as an increasing percentage of these outputs is rejected based on the uncertainty scores. Formally,

$$PRR = \frac{\text{AUC}_{\text{unc}} - \text{AUC}_{\text{rnd}}}{\text{AUC}_{\text{oracle}} - \text{AUC}_{\text{rnd}}}. \tag{3}$$

Rejecting 100% of the outputs is impractical, so we evaluate PRR at a 50% rejection rate, indicating how effectively the uncertainty score identifies and rejects the least desirable generations.

Quality metrics. For each evaluated task, we select a specific quality measure. For summarization, quality is assessed using the Align Score between input text and the output (Zha *et al.*, 2023). For translation, we employ COMET (Rei *et al.*, 2022). For short-form QA, we use Accuracy, while for long-form QA, we rely on the Align Score between the target answer and the generated output.

Baselines. We employ three simple token-probabilities-based baselines: Maximum Sequence Probability, Perplexity, and Mean Token Entropy (Fomicheva *et al.*, 2020). Additionally, we use state-of-the-art methods: Semantic Entropy (Kuhn *et al.*, 2023), SAR (Duan *et al.*, 2024), Degree Matrix and Eigenvalue of the Graph Laplacian (Lin *et al.*, 2023). We also include CoCoA variants and the Consistency score as defined in Vashurin *et al.* (2025).

Additional results. We extend the results from the main text and provide additional datasets in Tables 22, 23, and 24.

Signals of different types. The composition of Table 3 consists mostly of CoCoA variants. Table 21 uses signals of different types instead. Maximum Sequence Probability (MSP) is a probabilistic measure based on the log-probability of the greedy generation. Monte Carlo Normalized Sequence Entropy (MCNSE) estimates the length-normalized predictive entropy over several sampled generations. Greedy Average Dissimilarity (GreedyDissim) is a semantic measure, the average dissimilarity between the greedy generation and several sampled generations. The Eigenvalue Laplacian score (EigVal) is computed from the spectrum of the NLI entailment graph over sampled generations. The standard deviation is computed over 1000 bootstrap resamples of the evaluation objects, with the calibration split and the fitted aggregation kept fixed.

On Trivia and on average, VecUQ-OT is second only to the supervised model, which uses labels. This agrees with Tables 3 and 4, where VecUQ-OT also has the best average score among label-free methods. On MMLU, a single measure (MSP) performs best.

K Evaluation Details

For each image dataset, we split the 10,000 in-distribution test inputs into three disjoint parts with the same split for all ensemble groups. 10% of the inputs (1000, with labels) are used to fit the class

Method	Trivia	WMT19DeEn	MMLU	Avg
VecUQ-OT	0.6531 ± 0.0120	0.5678 ± 0.0131	0.4679 ± 0.0090	0.5629 ± 0.0066
Supervised model (uses labels)	**0.6612 ± 0.0117**	**0.6069 ± 0.0123**	0.5125 ± 0.0087	**0.5935 ± 0.0063**
Mean of component ranks	0.6506 ± 0.0120	0.5621 ± 0.0129	0.4604 ± 0.0090	0.5577 ± 0.0066
L2 norm after min-max	0.6498 ± 0.0120	0.5602 ± 0.0132	0.4607 ± 0.0090	0.5569 ± 0.0066
Whitened sum	0.5910 ± 0.0125	0.5776 ± 0.0132	0.4332 ± 0.0091	0.5339 ± 0.0068
MSP	0.6172 ± 0.0127	0.4538 ± 0.0149	**0.5143 ± 0.0087**	0.5285 ± 0.0072
MCNSE	0.5388 ± 0.0136	0.4446 ± 0.0151	0.3273 ± 0.0095	0.4369 ± 0.0073
GreedyDissim	0.6415 ± 0.0118	0.4791 ± 0.0153	0.4383 ± 0.0092	0.5196 ± 0.0069
EigVal	0.6257 ± 0.0124	0.3380 ± 0.0162	0.3893 ± 0.0094	0.4510 ± 0.0075

Table 21: PRR for a composition of signals of different types. Best in bold, second-best underlined.

Method	CoQA	GSM8k	MMLU	Trivia	WMT14FrEn	WMT19DeEn	XSum	mean
Ours (Exp, FW)	0.393	0.577	0.473	0.678	0.445	0.640	0.337	0.506
Ours (Uniform, FW)	0.395	0.568	0.472	0.680	0.441	0.643	0.339	0.505
CoCoA MSP	0.403	0.548	0.469	0.677	0.402	0.607	0.331	0.491
CoCoA PPL	0.389	0.459	0.469	0.678	0.396	0.571	0.309	0.467
CoCoA MTE	0.376	0.490	0.449	0.676	0.397	0.565	0.312	0.466
MSP	0.350	0.491	0.478	0.634	0.332	0.473	0.278	0.434
SAR	0.332	0.483	0.418	0.638	0.372	0.519	0.089	0.407
Consistency	0.403	0.465	0.427	0.652	0.306	0.499	0.063	0.402
PPL	0.289	0.280	0.478	0.636	0.370	0.484	0.226	0.395
MTE	0.252	0.326	0.459	0.623	0.397	0.477	0.203	0.391
DegMat	0.352	0.327	0.410	0.651	0.224	0.385	0.130	0.354
Semantic Entropy	0.294	0.483	0.387	0.565	0.274	0.409	-0.007	0.343
EigValLaplacian	0.316	0.264	0.401	0.622	0.201	0.330	0.126	0.323

Table 22: PRR (higher is better) for selective generation with Mistral-7B on all seven text datasets, extending Table 3. The top block shows VecUQ-OT with the exponential (Exp) or uniform reference and feature-wise (FW) scaling. As in Table 3, VecUQ-OT aggregates the measures of the second block (the CoCoA variants and MSP). The bottom block reports other uncertainty measures for reference. The last column is the mean over datasets.

means and the shared covariance of the Mahalanobis score for each ensemble member. 18% of the inputs (1800, without labels) form the calibration set, on which we fit the OT map of VecUQ-OT and the label-free aggregation baselines. The remaining 72% (7200) are used for evaluation, together with the OoD test inputs. Risk-based measures are computed from the predicted class distributions of the ensemble members. For OoD detection, the ROC-AUC is computed between the scores of in-distribution test inputs (negatives) and OoD inputs (positives). For misclassification detection,

Method	CoQA	GSM8k	MMLU	Trivia	WMT14FrEn	WMT19DeEn	XSum	mean
Ours (Exp, FW)	0.354	0.420	0.484	0.597	0.476	0.612	0.405	0.478
Ours (Uniform, FW)	0.356	0.409	0.481	0.599	0.476	0.613	0.410	0.478
CoCoA MSP	0.367	0.362	0.492	0.603	0.443	0.584	0.395	0.464
CoCoA PPL	0.352	0.417	0.458	0.598	0.440	0.509	0.401	0.453
CoCoA MTE	0.351	0.433	0.408	0.604	0.436	0.505	0.392	0.447
MSP	0.290	0.310	0.516	0.538	0.320	0.469	0.341	0.398
PPL	0.261	0.284	0.469	0.520	0.346	0.402	0.383	0.381
SAR	0.324	0.389	0.360	0.593	0.412	0.472	0.057	0.372
Consistency	0.391	0.367	0.395	0.615	0.376	0.450	0.007	0.372
MTE	0.248	0.308	0.362	0.506	0.358	0.391	0.367	0.363
DegMat	0.370	0.294	0.346	0.616	0.220	0.357	0.076	0.326
Semantic Entropy	0.289	0.384	0.235	0.541	0.277	0.410	0.029	0.309
EigValLaplacian	0.346	0.264	0.296	0.600	0.152	0.283	0.075	0.288

Table 23: PRR (higher is better) for selective generation with LLaMA-8B on all seven text datasets, extending Table 3. The top block shows VecUQ-OT with the exponential (Exp) or uniform reference and feature-wise (FW) scaling. As in Table 3, VecUQ-OT aggregates the measures of the second block (the CoCoA variants and MSP). The bottom block reports other uncertainty measures for reference. The last column is the mean over datasets.

Method	CoQA	GSM8k	MMLU	Trivia	WMT14FrEn	WMT19DeEn	XSum	mean
Ours (Exp, FW)	0.419	0.514	0.543	0.698	0.484	0.611	0.208	0.497
Ours (Uniform, FW)	0.423	0.506	0.543	0.698	0.486	0.615	0.210	0.497
CoCoA MTE	0.416	0.505	0.528	0.689	0.443	0.568	0.208	0.479
CoCoA MSP	0.410	0.438	0.539	0.699	0.444	0.590	0.222	0.477
CoCoA PPL	0.421	0.471	0.539	0.683	0.442	0.573	0.209	0.477
MSP	0.338	0.386	0.548	0.680	0.328	0.420	0.177	0.411
Consistency	0.417	0.407	0.493	0.657	0.332	0.487	0.203	0.428
MTE	0.305	0.384	0.543	0.642	0.426	0.532	0.139	0.424
SAR	0.392	0.373	0.520	0.647	0.397	0.503	0.136	0.424
PPL	0.323	0.343	0.548	0.653	0.394	0.520	0.144	0.418
DegMat	0.401	0.379	0.494	0.663	0.291	0.440	0.174	0.406
Semantic Entropy	0.306	0.420	0.474	0.593	0.323	0.411	0.149	0.383
EigValLaplacian	0.379	0.348	0.471	0.656	0.241	0.400	0.175	0.381

Table 24: PRR (higher is better) for selective generation with Falcon-7B on all seven text datasets, extending Table 3. The top block shows VecUQ-OT with the exponential (Exp) or uniform reference and feature-wise (FW) scaling. As in Table 3, VecUQ-OT aggregates the measures of the second block (the CoCoA variants and MSP). The bottom block reports other uncertainty measures for reference. The last column is the mean over datasets.

the positives are the test inputs on which the prediction of the ensemble is wrong. For selective prediction, we sort test inputs by increasing uncertainty, compute the accuracy on the first k inputs for each coverage level k/n, and report the area under the resulting accuracy-coverage curve.

L Additional Aggregation Experiments

Aggregation rules. All rules below are fitted on the calibration set and use the same uncertainty vectors as VecUQ-OT. *Standardized sum* computes the mean $\hat{\mu}_j$ and standard deviation $\hat{\sigma}_j$ of each component, applies the z-transform, and sums the coordinates. *Mean of component ranks* replaces each component with its empirical CDF and averages the results. *L2 norm after min-max* applies feature-wise min-max scaling, as in VecUQ-OT, and takes the L2 norm without an OT map. *Whitened sum* centers the vector with the calibration mean, whitens it with the regularized calibration covariance, and sums the coordinates. The *supervised model* requires labels. For images, we fit a logistic regression on binary labels of the target problem (OoD vs. InD, mistake vs. correct prediction) and obtain out-of-fold predictions with 4-fold cross-validation. Since it uses labeled samples from the same OoD dataset on which it is evaluated, it is an oracle-informed reference. For texts, we fit a ridge regression on the answer quality using 25% of the calibration split and evaluate on the remaining 75%. Table 4 reports the average results.

Redundant components. Fixed rules give a duplicated measure a larger weight, although it adds no information (see Appendix H for a related special case). Table 25 starts from the vector of Table 1 and adds nine exact copies of MahS, so that the vector has 13 components. All methods degrade, but VecUQ-OT degrades much less.

Task	#MahS	VecUQ-OT	Standardized sum	Mean of comp. ranks
OoD ROC-AUC	1	0.8739 ± 0.0020	0.8721 ± 0.0021	0.8615 ± 0.0028
OoD ROC-AUC	10	0.8327 ± 0.0099	0.7466 ± 0.0070	0.7197 ± 0.0078
Miscls. ROC-AUC	1	0.8817 ± 0.0021	0.8736 ± 0.0022	0.8757 ± 0.0026
Miscls. ROC-AUC	10	0.8108 ± 0.0077	0.7178 ± 0.0034	0.7016 ± 0.0040
Sel. pred. AUC	1	0.9335 ± 0.0005	0.9294 ± 0.0005	0.9323 ± 0.0006
Sel. pred. AUC	10	0.9059 ± 0.0027	0.8660 ± 0.0014	0.8561 ± 0.0013

Table 25: Effect of nine duplicated MahS components, averaged over datasets. Here, VecUQ-OT uses one outer anchor per axis.

Healthy and collapsed components. We split the nine pairs of Table 1 into two groups. For each pair, we compute the gap between the highest and the lowest ROC-AUC among the four individual components in Table 1. In the three CIFAR10 pairs, this gap is at most 0.03, so all components

work, and we call these pairs *healthy*. In the six other pairs, the gap is at least 0.19 because MahS collapses, and we call these pairs *collapsed*. Table 26 compares VecUQ-OT with the best component of each row, which is an oracle quantity because it requires labeled data of the task at hand. Table 27 compares VecUQ-OT with each component fixed a priori.

Group	Pairs	Ours	Best	Mean
Healthy	3	0.931	0.930	0.924
Collapsed	6	0.846	0.845	0.745

Table 26: Average OoD ROC-AUC of VecUQ-OT, of the best component of each row, and of all components.

Measure	Mean	Worst	Best
VecUQ-OT	0.874	0.752	1.000
$R^{1,2}_{\mathrm{tot}}$	0.863	0.735	0.963
R^{1}_{b}	0.848	0.724	0.963
$R^{1,2}_{\mathrm{exc}}$	0.846	0.737	1.000
MahS	0.661	0.405	0.934

Table 27: OoD ROC-AUC over the nine pairs of Table 1 for each measure fixed a priori.

Noisy components. A component that performs poorly is not the same as a noisy one. For example, MahS reaches ROC-AUC 0.405 on TinyImageNet vs. ImageNet-R, which is below 0.5 and thus still carries structure. To test truly uninformative components, we add j independent noise components to the vector of Table 1. Table 28 shows that the performance decreases as noise is added, while for $j = 5$ the noise components already outnumber the four informative ones.

Noisy components j	OoD ROC-AUC	Miscls. ROC-AUC
0	0.8740 ± 0.0020	0.8815 ± 0.0021
1	0.8520 ± 0.0024	0.8567 ± 0.0032
2	0.8361 ± 0.0029	0.8366 ± 0.0039
3	0.8231 ± 0.0033	0.8154 ± 0.0044
5	0.7952 ± 0.0042	0.7753 ± 0.0050

Table 28: VecUQ-OT with j added noise components, averaged over the pairs of Table 1.

M Runtime

We measure the fitting and inference time of VecUQ-OT without outer anchors and compare it with the standardized sum. The components are filled with samples from a log-normal$(0, 1)$ distribution, inference time is reported for 1000 vectors, and the standard deviation is computed over repeated runs on the same data. Table 29 fixes $N = 1800$, as in our experiments, and varies m. The fitting time barely changes because it is dominated by the $N \times N$ kernel, while the inference time grows with m. Table 30 fixes $m = 4$ and varies N. Fitting scales approximately quadratically with N, and inference approximately linearly. VecUQ-OT is fitted only once, and inference requires no optimization.

m	Fit (s)	Inference (ms)	Standardized fit (ms)	Standardized inference (ms)
1	1.261 ± 0.009	17.86 ± 0.11	0.042	0.012
5	1.270 ± 0.008	17.20 ± 0.25	0.111	0.038
10	1.241 ± 0.015	18.02 ± 0.14	0.148	0.048
50	1.258 ± 0.007	22.62 ± 0.11	0.268	0.106
100	1.268 ± 0.012	28.05 ± 0.14	0.420	0.179

Table 29: Runtime for $N = 1800$. The last two columns are the standardized sum.

N	Fit (s)	Inference (ms)	Standardized fit (ms)	Standardized inference (ms)
50	0.0040 ± 0.0002	0.661 ± 0.009	0.040	0.037
100	0.0061 ± 0.0001	1.074 ± 0.002	0.040	0.036
500	0.0679 ± 0.0006	4.337 ± 0.033	0.057	0.036
1000	0.3006 ± 0.0026	9.261 ± 0.049	0.075	0.035
5000	9.2891 ± 0.0112	62.669 ± 0.152	0.227	0.036
10000	35.6627 ± 0.0829	117.360 ± 0.639	0.412	0.036

Table 30: Runtime for $m = 4$. The last two columns are the standardized sum.

N Computation Resources

Our method is fully post-hoc, so it adds no training cost. To prepare pretrained ensembles and for LLM inference, we used NVidia GeForce 1080Ti GPU, while fitting VecUQ-OT and computing all metrics ran on CPU.

1. **Claims**

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

   Answer: [Yes]

   Justification: The Abstract and Section 1 state contributions that are matched by Sections 2, 3 and Section 5 (experiments).

2. **Limitations**

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

   Answer: [Yes]

   Justification: Yes, limitations are discussed in Appendix B.1, Section 3, and Appendix E.

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

   Justification: No new theorems are introduced. Properties used in Section 2.2 are cited from (Chernozhukov *et al.*, 2017; Hallin *et al.*, 2021).

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Yes. The code is provided. The datasets, models, metrics, and OT hyperparameters are specified in Section 5 and Appendices E, I, J.

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

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

   Answer: [Yes]

   Justification: The code is publicly available at `https://github.com/stat-ml/multidimensional_uncertainty`, with a README describing how to reproduce the experiments. All datasets are public and cited.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Reported in Section 5 and Appendices E, I, J, with ablations in Appendices C and F.

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Tables 1, 2, and Appendix I report mean and one-sigma standard deviation over four deep-ensemble groups of five members.

8. **Experiments compute resources**

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

   Answer: [Yes]

Justification: Yes, summarized in Appendix N.

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics.

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Foundational UQ methodology with positive impact on reliable decision-making and no specific negative path beyond generic UQ risks.

11. **Safeguards**

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

    Answer: [N/A]

    Justification: No high-risk models or datasets are released.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All datasets and pretrained models are cited.

13. **New assets**

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

    Answer: [Yes]

    Justification: An anonymized code package with a README is provided. No new datasets or models are introduced.

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

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

    Answer: [N/A]

    Justification: No crowdsourcing or human-subjects research.

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

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

    Answer: [N/A]

    Justification: No human-subjects research.

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs used for evaluation in the Section 5 are not part of the method.