WRSS: SEVERITY TRANSPORT FOR HIERARCHICAL RANKING EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Hierarchical classification errors differ not only in correctness, but also in severity. For ranked predictions, evaluation must additionally account for where severe alternatives appear. We study this problem as *severity displacement*: the movement of hierarchy-derived severity toward the head of a predicted ranking. We introduce the Wasserstein Rank–Severity Score (WRSS), which measures this displacement through one-dimensional transport over rank positions. The resulting distance is equivalent to a linear rank-weighted severity objective, while the transport construction jointly determines its weighting, normalization, equal-tier invariance, and mass-preserving top-k truncation. Controlled experiments separately test stability under stochastic ranking ambiguity and responsiveness to hierarchy-derived severity corruption. WRSS exhibits lower local volatility than HOPS in the tested ambiguity regimes and stronger alignment with controlled severity displacement in several fixed-budget settings. Experiments on trained classifiers show similar stability under input perturbations. WRSS therefore complements distance-based and preference-based metrics by measuring cumulative severity displacement along the ranking.

1 INTRODUCTION

Many recognition problems have labels organized by a taxonomy. In fine-grained species identification, aircraft recognition, and large-scale image classification, this structure captures semantic proximity, operational severity, or both. Errors are therefore not equally consequential: confusing a hawk with a falcon is generally less severe than confusing it with a fish. This motivates evaluation methods that measure not only whether a prediction is correct, but also how its errors relate to the label hierarchy. Such evaluation is inherently multifaceted, as different metrics encode different notions of hierarchical quality (Plaud et al., 2024; Kosmopoulos et al., 2015).

Distance-based metrics such as mistake severity and *average hierarchical distance* (AHD) evaluate how far predicted labels lie from the target in the hierarchy. They are interpretable and widely used, but their top-k variants typically summarize the retrieved labels without accounting for their internal order. Consequently, two rankings containing the same labels may receive the same score even when one places its most severe error first and the other places it last.

The *Hierarchically Ordered Preference Score* (HOPS) addresses this limitation by evaluating agreement between the predicted ranking and a hierarchy-induced preference ordering (Sani & Anand, 2026). This introduces rank sensitivity into hierarchical evaluation and captures information that distance-only summaries omit. It also defines a particular evaluation object: consistency with a weighted preference schedule derived from the hierarchy.

We study a complementary object, which we call *severity displacement*. Given a target class, the hierarchy assigns a severity to every candidate label. A predicted ranking then arranges these severities over rank positions. Severe labels near the head represent greater displacement from the hierarchy-optimal ordering than the same labels near the tail, while exchanging labels of equal severity should have no effect. This distinction separates three related evaluation questions: AHD measures hierarchical proximity irrespective of order, HOPS measures preference-order consistency, and severity displacement measures how strongly hierarchical error is concentrated toward early ranks.

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

We formulate severity displacement as one-dimensional transport along the ranking axis. The predicted ordering and the hierarchy-optimal ordering define two rank–severity profiles containing the same multiset of severity values. Their Wasserstein distance measures the cumulative work required to recover the optimal profile. We normalize this distance against the reversed severity profile to obtain the *Wasserstein Rank–Severity Score* (WRSS).

In this setting, the transport distance is equivalent to a linear rank-weighted severity objective. The transport formulation remains useful because it derives the rank weights, optimal reference, tight worst-case normalization, equal-tier invariance, and mass-preserving top-k treatment from one construction. Under truncation, severity outside the visible budget is aggregated into a tail bin, making permutations within the hidden tail irrelevant while preserving penalties for severity crossing the budget boundary.

We evaluate WRSS through two complementary criteria. First, stochastic ranking perturbations test local stability under ambiguity among similarly scored labels. Second, controlled hierarchy corruptions test responsiveness to known severity displacement. Considering both is necessary: stability alone may indicate insensitivity, while responsiveness without stability may indicate excessive volatility. We additionally test whether the observed stability extends to trained classifiers under input perturbations.

Contributions.

- We formulate hierarchical ranking evaluation as transport of severity over rank positions and introduce WRSS, a bounded score for cumulative rank–severity displacement.
- We show that the transport distance induces an equivalent linear rank-weighted severity objective together with a tight normalization, equal-tier invariance, and mass-preserving top-k truncation.
- We characterize WRSS through complementary tests of local stability and controlled severity responsiveness across dataset-derived hierarchies and trained-model outputs.

2 Related Work

Hierarchical evaluation metrics. Hierarchical classification evaluation has a long history. Surveys and unified treatments show that hierarchy structure changes what constitutes a "near" or "severe" error and document trade-offs among distance-, set-, and depth-aware measures (Costa et al., 2007; Kosmopoulos et al., 2015). More recent work in hierarchical and extreme multi-label classification shows that metric choice and inference procedure can materially alter conclusions under class imbalance and varying hierarchy depth (Amigó & Delgado, 2022; Plaud et al., 2024).

Severity and rank order. Mistake severity and average hierarchical distance score predictions by their distance from the target in the hierarchy. These measures are interpretable, but common top-k extensions aggregate the retrieved labels without preserving their internal order. They therefore cannot distinguish two top-k lists containing the same labels in different positions. The Hierarchically Ordered Preference Score (HOPS) addresses this limitation by comparing the predicted ranking with a hierarchy-induced preference ordering derived from LCA distances (Sani & Anand, 2026). HOPS evaluates weighted preference-order consistency, whereas WRSS evaluates cumulative displacement of hierarchy-derived severity over rank positions. The two metrics are rank-sensitive but use different weighting constructions.

Optimal transport and structured labels. Wasserstein distances measure the minimum work required to move mass between distributions and define a metric under a specified ground geometry (Villani, 2009; Santambrogio, 2015; Panaretos & Zemel, 2019). Existing uses of Wasserstein distance with structured labels generally transport probability mass in label space, with semantic or hierarchical distance defining the ground cost between classes (Frogner et al., 2015; Toyokuni et al., 2021). Related domain-specific methods likewise incorporate hierarchy into EMD-style losses or metrics (Yu et al., 2025). WRSS instead transports hierarchy-derived severity over a one-dimensional rank axis. The hierarchy determines the severity profile, while rank position determines the transport geometry. This separates WRSS from tree-Wasserstein distances between label distributions and makes the construction directly applicable to budgeted rankings with equal-severity tiers.

3 Problem Setup and Notation

Let $\mathcal{T}$ be a rooted hierarchy with leaf-label set $\mathcal{Y}$, where $|\mathcal{Y}| = K$. Given an input x, a model produces logits $s(x) \in \mathbb{R}^K$ or probabilities $p(x) \in \Delta^{K-1}$. Their descending order induces a ranking $\pi = (\pi(1), \ldots, \pi(K))$, where $\pi(j)$ is the label at rank j.

For a ground-truth leaf $y \in \mathcal{Y}$, let $d_{\mathcal{T}}(y, c)$ denote a hierarchy distance from y to $c \in \mathcal{Y}$. We write

$$r(y, c) \in \mathbb{R}_{\geq 0}$$

for the corresponding severity, with $r(y, y) = 0$ and larger values denoting more severe errors. The severity scale forms part of the evaluation geometry. Nonlinear choices of r encode different assumptions about the relative cost of distant errors.

Evaluation budgets. We consider two budgeted settings. Top-k evaluation exposes the first k ranks. Radius-d evaluation exposes the hierarchy-optimal labels satisfying $r(y, c) \leq d$, producing a target-dependent list length. We denote the corresponding metric variants by "@k" and "@d".

Equal-severity tiers. A hierarchy commonly assigns the same severity to many labels. For severity value t, define

$$n_t(y) = |\{c \in \mathcal{Y} : r(y, c) = t\}| .$$

These tier cardinalities may vary substantially across targets and severities. A rank-sensitive metric must therefore remain invariant to exchanges within an equal-severity tier while retaining sensitivity to severity moved across rank or budget boundaries.

4 Severity Transport: Wasserstein Rank–Severity Score

A hierarchy-aware ranking metric should distinguish both which severities occur and where they occur. A severe label near the head should contribute more than the same label near the tail, while exchanging labels of equal severity should have no effect. Under a finite budget, the metric should ignore order within the hidden tail but detect severity crossing into the visible ranking.

WRSS measures these properties through severity transport over rank positions. We first represent each ranking by its rank–severity profile, then compare the predicted profile with the hierarchy-optimal profile using one-dimensional Wasserstein distance.

Rank–severity profiles. Fix a target y and predicted ranking π. The hierarchy-optimal profile is the sorted severity multiset

$$z_j \coloneqq \operatorname{sort}\left(\{r(y, c) : c \in \mathcal{Y}\}\right)_j , \qquad j = 1, \ldots, K, \tag{1}$$

and the predicted profile is

$$\hat{z}_j \coloneqq r(y, \pi(j)) , \qquad j = 1, \ldots, K. \tag{2}$$

The two profiles contain the same severity multiset and differ only in its placement over rank.

Transport distance over rank. Let C be the prefix-sum operator

$$(Cx)_j = \sum_{i=1}^{j} x_i .$$

For nonnegative profiles x, z with equal total mass, define

$$W_1(x, z) \coloneqq \sum_{j=1}^{K} |(Cx)_j - (Cz)_j| . \tag{3}$$

This is the discrete one-dimensional Wasserstein-1 distance on a unit-spaced rank axis. Applied to $\hat{z}$ and z, it measures the cumulative work required to move severity from its predicted positions to its hierarchy-optimal positions.

The profiles need not be probability distributions. They define equal-mass finite measures because $\hat{z}$ and z are permutations of the same severity multiset.

Normalized score. Let

$$z^\star = (z_K, \dots, z_1)$$

be the reversed optimal profile. We define

$$\mathrm{WRSS}(\pi \mid y) \coloneqq 1 - \frac{W_1(\hat{z}, z)}{W_1(z^\star, z)}. \tag{4}$$

The reverse profile maximizes the transport distance over permutations of the severity multiset, up to permutations within equal-severity tiers. Consequently,

$$0 \le \mathrm{WRSS}(\pi \mid y) \le 1,$$

with 1 attained by a hierarchy-optimal ranking and 0 by the reversed severity ordering. Proofs are given in Appendix C.

The distance W_1 is the metric object. WRSS is a target-normalized score derived from that distance and is not itself a metric between arbitrary rankings. Additional metric properties of W_1 are established in Appendix D.

Top-k evaluation by tail lumping. Under top-k evaluation, order within ranks $k+1, \dots, K$ is unobserved, but severity crossing the top-k boundary remains relevant. We represent the hidden tail by its total severity:

$$T_k(x) \coloneqq \left(x_1, \dots, x_k, \sum_{i=k+1}^{K} x_i \right). \tag{5}$$

The budgeted score is

$$\mathrm{WRSS@k}(\pi \mid y) \coloneqq 1 - \frac{W_1(T_k(\hat{z}), T_k(z))}{W_1(T_k(z^\star), T_k(z))}. \tag{6}$$

Tail lumping preserves total severity mass and makes permutations within the hidden tail irrelevant. Severity moved between the visible prefix and hidden tail still changes the score.

Linear rank-weighted form. Because z is sorted in nondecreasing order, it minimizes every prefix sum among permutations of the same multiset. Hence,

$$(C\hat{z})_j - (Cz)_j \ge 0 \qquad \text{for all } j, \tag{7}$$

and the absolute values in Equation (3) may be removed:

$$\begin{aligned} W_1(\hat{z}, z) &= \sum_{j=1}^{K} \left((C\hat{z})_j - (Cz)_j \right) \\ &= \sum_{i=1}^{K} (K - i + 1)\hat{z}_i - \sum_{i=1}^{K} (K - i + 1) z_i. \end{aligned} \tag{8}$$

Thus, relative to the fixed optimal-profile term and reversed-profile normalization, WRSS is a linear rank-weighted severity score. The weights $K, K-1, \dots, 1$ follow from unit-distance transport over rank rather than being specified separately.

After tail lumping, the visible ranks receive weights $k, k-1, \dots, 1$ while the equal-mass tail bin contributes no positional term. Given a ranking, WRSS is evaluated in $O(K)$ time without an iterative transport solver; sorting adds $O(K \log K)$ when required.

Radius-d evaluation. A radius budget fixes semantic tolerance rather than list length. Define

$$k_d(y) \coloneqq \sum_{j=1}^{K} \mathbf{1}\{z_j \le d\}, \tag{9}$$

the number of hierarchy-optimal labels within severity radius d. We then set

$$\mathrm{WRSS@d}(\pi \mid y) \coloneqq \mathrm{WRSS@k}(\pi \mid y)\big|_{k = k_d(y)}. \tag{10}$$

Fixed-k evaluation holds list length constant. Radius-d evaluation holds semantic tolerance constant and allows the corresponding list length to vary with the local hierarchy.

Properties. WRSS has four immediate properties. First, larger severities incur larger costs through the values $r(y, c)$. Second, moving severity toward earlier ranks changes more cumulative prefixes and therefore incurs a larger penalty. Third, tail lumping preserves the top-k boundary while removing dependence on order within the hidden tail. Fourth, exchanging equal-severity labels leaves the rank–severity profile unchanged.

Hard rankings remain piecewise constant functions of model logits. WRSS does not remove that discontinuity. Instead, when a perturbation changes the ranking, the magnitude of the score change depends on the induced severity displacement.

Severity-scale behavior. WRSS is invariant to positive affine transformations of severity. Let

$$r'(y, c) = a\, r(y, c) + b, \qquad a > 0.$$

The corresponding profiles satisfy

$$\hat{z}' = a\hat{z} + b\mathbf{1}, \qquad z' = az + b\mathbf{1}.$$

The additive terms cancel in every prefix difference, while both the numerator and denominator in Equation (4) scale by a. The normalized score is therefore unchanged.

Nonlinear transformations, such as r^2 or $\log(1 + r)$, alter the relative cost of severity levels and may change the score. They correspond to different cardinal assumptions about distant errors rather than equivalent parameterizations of the same geometry.

5 BASELINES

We compare WRSS with a distance-based metric and a preference-based metric. These methods evaluate related but distinct properties of hierarchical predictions.

Average Hierarchical Distance. Average Hierarchical Distance (AHD) measures the mean hierarchy distance among the first k predictions:

$$\mathrm{AHD@k}(y, \pi) \coloneqq \frac{1}{k} \sum_{j=1}^{k} d_{\mathcal{T}}(y, \pi(j)) \,. \tag{11}$$

AHD directly summarizes hierarchical proximity and provides a natural reference for distance-derived oracle costs. It depends only on the multiset of distances within the top-k predictions and is therefore unchanged by their internal permutation.

Hierarchically Ordered Preference Score. HOPS measures agreement between the predicted ranking and a hierarchy-induced preference ordering. Let z be the sorted optimal tier profile and $\hat{z}_i = r(y, \pi(i))$ the predicted profile. Its weighted deviation is

$$N_{\mathrm{HOPS}}(\hat{z}, z) = \sum_{i=1}^{K} \eta_i \, |\hat{z}_i - z_i| \,, \tag{12}$$

and the normalized score is

$$\mathrm{HOPS}(\pi \mid y) = \mathrm{clip}_{[0,1]} \left(1 - \frac{\sum_{i=1}^{K} \eta_i |\hat{z}_i - z_i|}{\sum_{i=1}^{K} \eta_i |z_i^{\mathrm{worst}} - z_i|} \right). \tag{13}$$

For top-k evaluation, HOPS uses

$$z^{\mathrm{worst}} = (z_k, z_{k-1}, \ldots, z_1, z_{k+1}, \ldots, z_K), \tag{14}$$

with $\eta_i = 0$ for $i > k$. The full-list variant uses

$$z^{\mathrm{worst}} = (z_K, \ldots, z_1).$$

In our implementation, a contiguous block with optimal tier value u receives weights linearly interpolated from 2^{-u} to $2^{-(u+1)}$. Radius-d evaluation uses the induced budget $k_d(y)$ from Equation (9). At $k = 1$, HOPS equals top-1 correctness.

Relation between the metrics. AHD, HOPS, and WRSS differ primarily in the quantities they preserve. AHD aggregates distance without retaining rank order. HOPS evaluates deviation using a hierarchy-dependent preference weighting. WRSS evaluates cumulative severity displacement using the rank weights induced by one-dimensional transport.

For example, consider

$$z = (0, 1, 2, 3, 4)$$

and the predicted profiles

$$\hat{z}^{(1)} = (2, 1, 0, 4, 3), \qquad \hat{z}^{(2)} = (4, 1, 2, 3, 0),$$

under $k = 3$. Both HOPS and WRSS can distinguish these rankings. HOPS does so through its η-weighted deviation, whereas WRSS does so through the cumulative transport induced by placing severity 4 at the first rank. The comparison is therefore between different rank-weighting constructions, not between weighted and unweighted ranking errors.

6 Metric Stress-Test Battery

We use model-free stress tests to isolate the effects of hierarchy structure, rank position, and score ambiguity. The samplers are not intended to reproduce a particular predictor distribution. Instead, each sampler controls one property relevant to hierarchical ranking evaluation. Formal definitions appear in Appendix E.

Hierarchies. We use the dataset taxonomies of CIFAR-100, FGVC-Aircraft, and iNat19. These hierarchies differ in depth, branching structure, and equal-severity tier sizes, providing distinct settings for budgeted evaluation.

Tier-softmax sampler S_{TS}. The tier-softmax sampler produces rankings that generally follow hierarchy severity while retaining local ambiguity. Logits decrease with severity and receive spread-normalized Gumbel noise. The resulting rankings contain near-ties among hierarchy-proximal labels. This sampler tests local sensitivity when rankings are broadly hierarchy-consistent but uncertain.

Within-tier sampler S_{WT}. The within-tier sampler preserves the ordering between severity tiers and randomly permutes labels within each equal-severity tier. Optional post-ranking noise introduces small score perturbations without changing the intended tier structure. Because within-tier exchanges do not alter the severity profile, this sampler tests whether a metric responds to distinctions that are irrelevant under the specified hierarchy.

Head-corruption sampler S_{HC}. The head-corruption sampler swaps early-ranked labels with labels from distant severity tiers. The multiset of logits is preserved, but severe labels are moved toward the head. This sampler tests whether a metric penalizes severity according to its rank position.

Boundary corruption. Boundary corruption provides a controlled severity signal. For target y, we select an off-path subtree and move probability mass α to leaves within that subtree. We define

$$D_{\mathrm{root}}$$

as the distance from y to the selected subtree root, and

$$D_{\mathrm{leaf}}$$

as the mean distance from y to the leaves within that subtree. The corresponding oracle costs are

$$C_{\mathrm{root}} = \alpha D_{\mathrm{root}}, \qquad C_{\mathrm{leaf}} = \alpha D_{\mathrm{leaf}}. \tag{15}$$

These costs combine displaced mass with hierarchy distance; they specify the tested severity geometry rather than external semantic or operational harm.

7 Evaluation Criteria

We evaluate two complementary properties: stability under local perturbations and responsiveness to controlled severity changes. We additionally apply the same metrics to trained-model outputs. No single diagnostic defines a global ordering of AHD, HOPS, and WRSS because the metrics preserve different information.

Local stability under logit jitter. For fixed sampled logits s, we generate

$$\tilde{s} = s + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I),$$

using

$$\sigma \in \{0.0, 0.01, 0.02, 0.05, 0.1\}.$$

For each target leaf ℓ, we compute the maximum adjacent finite-difference slope

$$S_\ell(M) = \max_j \frac{\left|\bar{M}_{\ell,\sigma_{j+1}} - \bar{M}_{\ell,\sigma_j}\right|}{\sigma_{j+1} - \sigma_j}, \tag{16}$$

where $\bar{M}_{\ell,\sigma_j}$ is the mean over repeated jitter samples at noise level σ_j. We summarize

$$\log_{10}(S_\ell(M) + \epsilon)$$

using the median, interquartile range, and upper quantiles across targets and budget sweeps. Complete definitions are given in Appendix G.

Controlled severity separability. For boundary corruptions, the oracle costs C_{root} and C_{leaf} are defined in Equation (15). We use them to test whether metric values distinguish distinguish corruptions with lower and higher hierarchy-derived oracle cost.

Before computing near–far AUROC, all metric outputs are oriented as severity scores:

$$1 - \text{WRSS}, \qquad 1 - \text{HOPS}, \qquad \text{AHD}\,.$$

We also report signed rank correlation or monotone cost–score alignment where appropriate. Formal definitions are provided in Appendix F.

Evaluation on trained models. We evaluate leaf rankings from cross-entropy and Hier-COS models on CIFAR-100, FGVC-Aircraft, and iNat19. All outputs use the same AHD, HOPS, and WRSS implementations as the controlled experiments.

For input-jitter experiments, lightweight image perturbations are applied before repeated forward passes. We then compute aggregate local Lipschitz-proxy summaries over matched models, datasets, and budgets. These experiments test whether the local behavior observed under synthetic logit perturbations also appears for trained predictors.

For naturally occurring errors, the displaced mass α is unknown. We therefore use the distance-only quantities

$$C_{\text{root}} = D_{\text{root}}, \qquad C_{\text{leaf}} = D_{\text{leaf}}.$$

Because these reference quantities are hierarchy distances, AHD provides a direct distance-based baseline, while HOPS and WRSS retain rank-sensitive information.

8 Results

We evaluate two complementary properties. Local perturbation experiments measure stability under ranking ambiguity, while controlled corruptions measure sensitivity to hierarchy-derived severity. Stability alone is insufficient: a useful metric must remain responsive when severe labels move toward earlier ranks.

8.1 Stability under stochastic ranking ambiguity

Table 1 reports local sensitivity under controlled logit jitter on the iNat19 hierarchy. WRSS has lower median and P90 sensitivity than HOPS in all three regimes.

Because

$$\Delta_{\log_{10}} = Z_{\text{WRSS}} - Z_{\text{HOPS}},$$

the corresponding sensitivity ratio is

$$\frac{S_{\text{WRSS}}}{S_{\text{HOPS}}} = 10^{\Delta_{\log_{10}}}.$$

Thus, negative values indicate lower WRSS sensitivity. For example, $\Delta_{\log_{10}} = -2$ corresponds to a $100\times$ reduction. These results establish stability under local score perturbations. The next experiment tests whether WRSS remains responsive to controlled changes in hierarchical severity.

Table 1: **Local stability under logit jitter** (lower is better, ↓). Entries report the median [IQR] and P90 of $Z_\ell(M) = \log_{10}(S_\ell(M) + \epsilon)$, where $S_\ell(M)$ is the sample-wise local sensitivity defined in Section 7. Results are aggregated over leaves and budget sweeps on the iNat19 hierarchy. $\Delta_{\log_{10}}$ denotes WRSS minus HOPS.

	Median [IQR] log-sensitivity			**P90 log-sensitivity**		
Regime	WRSS	HOPS	$\Delta_{\log_{10}}$	WRSS	HOPS	$\Delta_{\log_{10}}$
S_{HC}	$-1.44\,[1.07]$	$-0.52\,[1.16]$	-0.92	-0.72	0.50	-1.22
S_{TS}	$0.23\,[0.92]$	$0.99\,[0.84]$	-0.76	1.06	1.66	-0.61
S_{WT}	$-3.80\,[8.93]$	$-2.30\,[10.71]$	-1.49	-2.39	-0.62	-1.77

8.2 Severity-displacement separability

We next test whether the metrics distinguish near and far hierarchy errors. The controlled corruptions define two oracle costs: C_{root}, based on distance between subtree roots, and C_{leaf}, which additionally resolves variation within the incorrect subtree. All metric outputs are oriented as severity scores before computing near–far AUROC.

Table 2 shows that WRSS separates near and far errors more reliably than HOPS for both oracles, except for the tied FGVC-Aircraft root-level result. The largest gains occur under C_{leaf}, where the oracle distinguishes severity within an incorrect subtree. Under C_{root}, this information is removed and the differences are smaller.

AHD is strongest on FGVC-Aircraft and competitive elsewhere. This is expected: the oracle costs are themselves hierarchy distances, and AHD directly averages such distances. WRSS answers a different question by retaining rank position. It measures whether severity is concentrated near the head of the list, whereas AHD is unchanged by permutations of the same predictions. The relevant rank-sensitive comparison is therefore between WRSS and HOPS.

Budget dependence. The fixed-k results in Appendix A show that WRSS gains are generally larger when more of the rank–severity profile is observed, particularly under C_{leaf}. Results under adaptive radius budgets are less uniform. Increasing d changes both the number of retained labels and their severity distribution, so scores at different radii do not isolate rank budget alone.

8.3 Evaluation on trained classifiers

Input-jitter stability. For trained models, we use the aggregate LLIP procedure from Section 7. Table 3 reports $\log_{10}$ sensitivity for cross-entropy and Hier-COS models on CIFAR-100, FGVC-Aircraft, and iNat19. WRSS is less sensitive than HOPS across the evaluated datasets, models, and both fixed-k and radius-d budgets. The controlled and trained-model experiments therefore agree that WRSS varies less under local perturbations of the predicted ranking.

Table 2: **Near–far severity separability** (AUROC, higher is better, ↑). WRSS and HOPS are converted to severity scores using $1 - \mathrm{score}$; AHD is already distance-oriented. Near and far samples are defined by oracle-cost quantiles. AHD provides a distance-only reference. Values are mean ± standard deviation over all defined experimental configurations, grouped by the displayed dataset, metric, and oracle. Values are rounded to three decimal places; a reported standard deviation of 0.000 indicates a value below 0.0005.

Metric	$AUROC(C_{\text{leaf}})$ ↑			$AUROC(C_{\text{root}})$ ↑		
Dataset	CIFAR-100	FGVC-Aircraft	iNat19	CIFAR-100	FGVC-Aircraft	iNat19
WRSS	$\mathbf{0.77} \pm 0.001$	0.70 ± 0.025	$\mathbf{0.66} \pm 0.005$	$\mathbf{0.68} \pm 0.002$	0.53 ± 0.001	$\mathbf{0.60} \pm 0.006$
HOPS	0.65 ± 0.000	0.62 ± 0.010	0.54 ± 0.002	0.60 ± 0.001	0.53 ± 0.002	0.51 ± 0.001
AHD	0.69 ± 0.001	$\mathbf{0.76} \pm 0.006$	0.62 ± 0.003	0.63 ± 0.001	$\mathbf{0.55} \pm 0.002$	0.59 ± 0.005

Table 3: **Real-model robustness under input jitter** (log_{10} Lipschitz-proxy ↓ (LLIP) lower is better, ↓).

		Cross-Entropy			Hier-COS		
Dataset	**Budget**	LLIP(WRSS)	LLIP(HOPS)	$\frac{\text{WRSS}}{\text{HOPS}}$ ↓	LLIP(WRSS)	LLIP(HOPS)	$\frac{\text{WRSS}}{\text{HOPS}}$ ↓
CIFAR	k	-2.40 ± 0.01	-1.84 ± 0.01	0.2769	-2.41 ± 0.01	-1.85 ± 0.01	0.2729
CIFAR	d	-2.62 ± 0.04	-1.86 ± 0.01	0.1728	-2.66 ± 0.01	-1.86 ± 0.01	0.1563
FGVC	k	-3.09 ± 0.06	-2.78 ± 0.08	0.4881	-3.02 ± 0.05	-2.76 ± 0.05	0.5444
FGVC	d	-3.09 ± 0.07	-2.68 ± 0.08	0.3914	-3.04 ± 0.06	-2.64 ± 0.05	0.4035
iNat19	k	-2.56 ± 0.02	-1.57 ± 0.02	0.1018	-2.51 ± 0.01	-1.49 ± 0.01	0.0961
iNat19	d	-2.92 ± 0.02	-1.67 ± 0.02	0.0561	-2.76 ± 0.01	-1.54 ± 0.01	0.0591

Severity separability on natural errors. Appendix B reports results on naturally occurring model errors. AHD is often strongest against the distance-derived oracles, while WRSS is competitive with HOPS and performs better in several deeper-hierarchy settings. These results are consistent with the metrics measuring related but distinct properties: AHD measures order-invariant severity, HOPS measures preference consistency, and WRSS measures severity displacement over rank.

9 Limitations

WRSS depends on the chosen hierarchy severity scale. We use LCA-derived severity tiers because they are standard and interpretable, but they represent only one possible model of semantic or operational cost. Normalized WRSS is invariant to positive affine transformations $r' = ar + b$, with $a > 0$: the additive term cancels between profiles, while the multiplicative factor scales the numerator and reversed-profile denominator equally. Nonlinear transformations such as r^2 or $\log(1 + r)$, as well as edge-weighted trees and task-specific costs, impose different cardinal assumptions about severe mistakes and may therefore change quantitative conclusions.

The normalization further depends on the chosen budgeted interface, although the underlying transport distance remains the central geometric quantity. Our stress-test samplers isolate metric behavior under controlled ranking ambiguities and corruptions, but should not be interpreted as complete models of real classifier errors. Finally, WRSS complements rather than universally replaces existing hierarchical metrics: AHD remains appropriate for measuring distance-only severity, while HOPS measures weighted preference-order consistency.

10 Conclusion

Hierarchical top-k evaluation must account for both the severity of mistakes and their positions in the ranking. We introduced the Wasserstein Rank–Severity Score (WRSS), which represents hierarchy-derived severities as mass on the rank axis and measures the one-dimensional transport required to recover a hierarchy-optimal profile. Under this construction, WRSS admits an equivalent normalized linear rank-weighted severity form, while the transport formulation jointly determines the rank weights, reversed-profile normalization, and top-k tail treatment.

Across controlled stress tests and dataset-derived hierarchies, WRSS is stable under small stochastic ranking perturbations while distinguishing rankings that differ in the displacement of severe errors. These results support a complementary metric-suite view of hierarchical evaluation: AHD measures distance-only severity, HOPS measures weighted preference-order consistency, and WRSS measures transport-derived rank–severity displacement. More broadly, severity transport provides a principled basis for budgeted hierarchical ranking evaluation and motivates future work on alternative severity geometries and differentiable hierarchy-aware objectives.

AI Use Statement

Generative AI tools were used to assist with manuscript drafting and editing, software development and debugging, methodological critique, and interpretation of experimental results. All AI-assisted

text, code, analyses, and claims were reviewed and verified by the authors. The authors determined the research questions, mathematical formulation, experimental design, and final claims, and take responsibility for the complete contents of this submission.

Reproducibility Statement

The definition, normalization, and computational properties of WRSS are provided in the main text, with additional derivations and proofs in the appendices. The experimental protocols, hierarchy construction, corruption procedures, oracle costs, perturbation settings, aggregation procedures, and hyperparameters are documented in the experimental sections and appendices. An anonymized code repository containing the implementation, experiment configurations, and analysis scripts is included in the supplementary material.

Ethics Statement

This work proposes an evaluation metric and uses publicly available benchmark datasets. It involves no human-subject data collection or deployment study. The hierarchy-derived severity values used by WRSS encode the supplied cost structure and should not be interpreted as independently validated measures of semantic, societal, or application-level harm.

References

Enrique Amigó and Agustín Delgado. Evaluating extreme hierarchical multi-label classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5809–5819, 2022. doi: 10.18653/v1/2022.acl-long.399. URL https://aclanthology.org/2022.acl-long.399/.

Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making better mistakes: Leveraging class hierarchies with deep networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12503–12512, 2019. URL https://api.semanticscholar.org/CorpusID:209415075.

Eduardo P. Costa, Ana C. Lorena, André C. P. L. F. Carvalho, and Alex A. Freitas. A review of performance evaluation measures for hierarchical classifiers. In *AAAI 2007 Workshop on Evaluation Methods for Machine Learning II*, 2007. URL https://cdn.aaai.org/Workshops/2007/WS-07-05/WS07-05-001.pdf.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. URL https://arxiv.org/abs/1506.05439.

Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. Evaluation measures for hierarchical classification: a unified view and novel approaches. *Data Mining and Knowledge Discovery*, 29(3):820–865, 2015. doi: 10.1007/s10618-014-0382-x. URL https://link.springer.com/article/10.1007/s10618-014-0382-x.

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

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Victor M Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6(1):405–431, 2019.

Roman Plaud, Matthieu Labeau, Antoine Saillenfest, and Thomas Bonald. Revisiting hierarchical text classification: Inference and metrics. In *Proceedings of the 28th Conference on Computational Natural Language Learning (CoNLL)*, pp. 231–242, 2024. doi: 10.18653/v1/2024.conll-1.18. URL https://aclanthology.org/2024.conll-1.18/.

Depanshu Sani and Saket Anand. Hier-cos: Making deep features hierarchy-aware via composition of orthogonal subspaces, June 2026.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Birkhäuser, 2015. doi: 10.1007/978-3-319-20828-2. URL https://link.springer.com/book/10.1007/978-3-319-20828-2.

Ayato Toyokuni, Sho Yokoi, Hisashi Kashima, and Makoto Yamada. Computationally efficient Wasserstein loss for structured labels. In *Proceedings of the EACL Student Research Workshop*, 2021. URL https://aclanthology.org/2021.eacl-srw.1/.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset-supplementary material. *Reptilia*, 32(400):1–3, 2017.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der Mathematischen Wissenschaften*. Springer, 2009. doi: 10.1007/978-3-540-71050-9. URL https://link.springer.com/book/10.1007/978-3-540-71050-9.

Hai-Tao Yu, Dawei Li, and Xin Kang. Emotional earth mover's distance for fine-grained hierarchical emotion analysis. In *Advanced Data Mining and Applications (ADMA)*, 2025. URL https://link.springer.com/chapter/10.1007/978-981-95-3453-1_20.