BEYOND LOGIT ADJUSTMENT: A Residual Decomposition Framework for Long-Tailed Reranking

Zhanliang Wang[1,2], **Hongzhuo Chen**[1,2], **Quan Minh Nguyen**[1,2], **Mian Umair Ahsan**[2], **Kai Wang**[1,2]*

[1] University of Pennsylvania, Philadelphia, PA, USA
[2] Children's Hospital of Philadelphia, Philadelphia, PA, USA

Abstract

Long-tailed classification, where a small number of frequent classes dominate many rare ones, remains challenging because models systematically favor frequent classes at inference time. Existing post-hoc methods such as logit adjustment address this by adding a fixed classwise offset to the base-model logits. However, the correction required to restore the relative ranking of two classes need not be constant across inputs, and a fixed offset cannot adapt to such variation. We study this problem through Bayes-optimal reranking on a base-model top-k shortlist. The gap between the optimal score and the base score, the *residual correction*, decomposes into a *classwise* component that is constant within each class, and a *pairwise* component that depends on the input and competing labels. When the residual is purely classwise, a fixed offset suffices to recover the Bayes-optimal ordering. We further show that when the same label pair induces incompatible ordering constraints across contexts, no fixed offset can achieve this recovery. This decomposition leads to testable predictions regarding when pairwise correction can improve performance and when cannot. We develop REPAIR (**Re**ranking via **Pair**wise residual correction), a lightweight post-hoc reranker that combines a shrinkage-stabilized classwise term with a linear pairwise term driven by competition features on the shortlist. Experiments across nine benchmarks spanning text classification, visual recognition, and multimodal rare-disease prediction confirm that the decomposition explains when pairwise correction is beneficial and when classwise correction alone suffices. Together, these results establish residual decomposition as a principled framework for understanding and improving post-hoc correction in long-tailed classification.

Code: https://github.com/WGLab/REPAIR

1 Introduction

Long-tailed classification, where a few common classes dominate many rare ones, remains a persistent challenge because models systematically favor frequent classes at prediction time, pushing rare labels down the ranking even when they are plausible. A common post-hoc correction method is logit adjustment (Menon et al., 2021), which adds a fixed class-dependent offset to the base-model scores. This is effective when rare classes are globally under-ranked due to bias introduced by training frequency, but long-tailed ranking errors are more complicated than a single classwise offset can express.

Consider an animal classifier trained on far more images of *cat* than *leopard* (Figure 1). On a clear photograph showing a leopard's rosette pattern, the model assigns 56% to cat and 44% to leopard. On a distant, partially occluded shot of a leopard, the model gives 70% to cat and only 30%to leopard. On a third image that genuinely depicts a cat, the model gives

*Corresponding author. Email: wangk@chop.edu

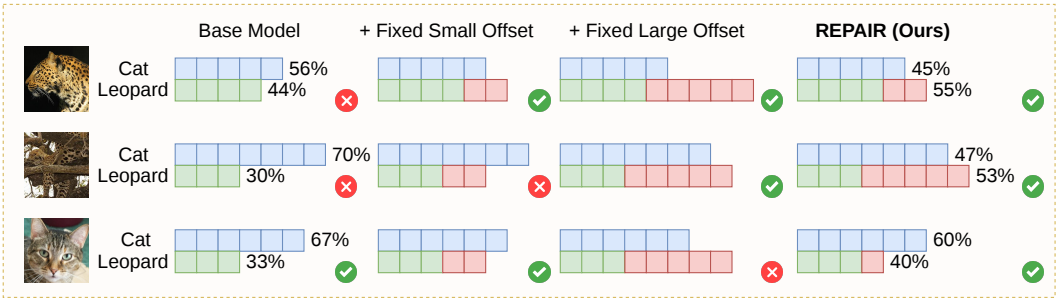


Figure 1: Motivating example (shortlist size 2 for clarity). Each row is a different input; bars show scores and percentages indicate softmax probabilities. A fixed small offset corrects the easy case (row 1) but under-corrects the hard case (row 2); a fixed large offset corrects both but flips the already-correct case (row 3). REPAIR adapts the correction magnitude per input, producing the correct ranking in all three cases.

67% to cat and 33% to leopard. No single fixed per-class offset works for all three: a value small enough to preserve the correct ranking on the third image cannot close the gap on the second, while a value large enough to correct the second inevitably flips the third. The root cause is data imbalance: with few training images of leopard, the base model's scores for this class are not only biased but highly variable across inputs, and this variability is exactly what a fixed offset cannot absorb.

We formalize this by studying Bayes-optimal reranking on a base-model shortlist. The gap between the optimal score and the base score, the *residual correction*, decomposes into a *classwise* component, which depends only on the label identity; and a pairwise component, which varies with the input and the competing labels; Section 4 characterizes precisely when the classwise component alone suffices and when it does not. Real datasets fall on a spectrum between these extremes, and the degree of pairwise competition determines how much a context-dependent correction can help. This raises a practical question: *how should a reranker adapt its correction to the specific input and the competing labels on the shortlist?* Our key insight is that the competition structure within the shortlist already encodes this information: how closely the base model scores the true label relative to each rival reflects how hard the current input is, and this relative strength varies across inputs, which is exactly the variation that a fixed offset misses.

Based on this analysis, we develop REPAIR (Reranking via Pairwise residual correction), a lightweight post-hoc reranker that combines a shrinkage-stabilized classwise term, which alone improves over logit adjustment by learning offsets on the shortlist, with a linear pairwise term driven by competition features. Both components are fitted jointly on covered held-out calibration examples; the base model is never modified. We validate on nine long-tailed benchmarks spanning text classification, visual recognition, and rare-disease diagnosis. Consistent with the decomposition, the pairwise term yields small gains where the classwise correction already suffices, as on the vision benchmarks (ImageNet-LT (Liu et al., 2019), Places-LT (Liu et al., 2019), iNaturalist (Van Horn et al., 2018)) and the class-separable text task GoEmotions (Demszky et al., 2020). Gains are substantially larger where it does not, as on the non-class-separable text tasks (MASSIVE (FitzGerald et al., 2023), HuffPost-LT (Misra, 2022), Few-NERD (Ding et al., 2021)) and rare-disease diagnosis on GMDB (Lesmann et al., 2024) and its out-of-distribution variant RareBench (Chen et al., 2024)).

Contributions.

1. We decompose the residual correction on a fixed shortlist into classwise and pairwise components, proving that when the same label pair induces incompatible ordering constraints across inputs, no fixed per-class offset can recover the Bayes ordering (Theorem 4.3).

2. We derive REPAIR, a lightweight post-hoc reranker from this decomposition, combining a learned classwise correction with a linear pairwise term.
3. We validate the framework's predictions on nine benchmarks: gains are proportional to the degree of pairwise competition and concentrated where the theory predicts fixed offsets to be the most fragile.

2 Related Work

Long-tailed distribution. Real world class distributions are rarely uniform. Typically a few head classes dominate many rare ones, forming a long-tailed distribution (Yang et al., 2022; Zhang et al., 2023). Models trained on such datasets are systematically biased towards frequent classes, which usually results in poor performance on the tail (Ren et al., 2020; Menon et al., 2021). Existing solutions can be grouped into two main categories based on when the correction is applied: training-time methods and post-hoc methods.

Post-hoc score correction. Post-hoc methods adjust the output of a fixed pretrained model at inference time. Logit adjustment methods add a fixed per-class offset to correct the distribution shift (Ren et al., 2020; Menon et al., 2021; Hong et al., 2021). Weight normalization methods rescale the classifier weight norms to remove the correlation between norm magnitude and class frequency (Kang et al., 2020; Kim & Kim, 2020). The correction for each class remains fixed, which is a common limitation shared by all these methods. Temperature scaling (Guo et al., 2017; Xiao et al., 2025) is also post-hoc but rank-preserving, targeting calibration rather than ordering. A broader review of training-time methods is provided in Appendix B.

Rare-disease diagnosis. Many rare genetic disorders present distinctive facial features, and CNN-based systems (Gurovich et al., 2019; Hsieh et al., 2022) can prioritize candidate diagnoses from facial photos across hundreds of syndromes. Recent multimodal approaches further integrate clinical notes and demographic information via large language models (Wu et al., 2026; Wang et al., 2026a;b), yet the extreme class imbalance inherent in rare disease settings remains a core challenge across all these systems.

3 Preliminaries

We consider a multiclass classification problem with input X, label $Y \in [K] = \{1, \ldots, K\}$, and long-tailed class prior $\pi_y = P(Y = y)$. A trained base model produces scores $g(X) = (g_1(X), \ldots, g_K(X)) \in \mathbb{R}^K$.

Shortlist and coverage. Fix an integer $k \ll K$ and let $S(X) \subseteq [K]$ denote the deterministic top-k shortlist under $g(X)$, with deterministic tie-breaking. A *reranker* selects a label $\hat{y}(X) \in S(X)$. Write $C := \{Y \in S(X)\}$ for the *coverage event*; every reranker satisfies $P(\hat{y}(X) = Y) \leq P(C) = \text{Recall@}k$, so shortlist recall is a hard ceiling on reranking accuracy. Since reranking can only affect covered examples, we condition all subsequent analysis and evaluation metrics on C.

Shortlist posterior. Once the shortlist is fixed, the relevant object for reranking is the posterior restricted to the shortlisted labels:

$$p_S(y \mid x) := P(Y = y \mid X = x, C), \qquad y \in S. \tag{1}$$

The corresponding log-posterior,

$$s^\star_y(x, S) := \log p_S(y \mid x), \tag{2}$$

is the Bayes-optimal shortlist score: a standard pointwise argument (Appendix D.4.1) shows that the Bayes-optimal reranker selects $\hat{y}^\star(x) \in \arg\max_{y \in S(x)} p_S(y \mid x)$. Throughout, only relative values of $s^\star_y$ across $y \in S$ matter for ranking. A generic classwise reranker modifies

the base score as $g_y(x) + a_y$ for a fixed per-class offset a_y; prior-based logit adjustment is the special choice $a_y = -\tau \log \pi_y$. The next section discusses when this classwise form is sufficient and when it is not.

4 A Residual Correction Framework for Long-Tailed Reranking

Once a base model returns a fixed shortlist, the first question is what score a Bayes-optimal reranker should target. The second is how the base score differs from this target. That gap defines a *residual correction*. We then ask when this correction can be represented by a fixed classwise term, and when context-dependent competition makes a pairwise term necessary.

4.1 Residual correction and pairwise gaps

For each shortlisted label $y \in S$, define the *residual correction*

$$\beta^\star_y(x, S) \;:=\; s^\star_y(x, S) \;-\; g_y(x). \tag{3}$$

Because ranking depends only on score *differences*, the relevant object is the *pairwise gap*

$$\Delta_{uv}(x, S) \;:=\; \beta^\star_u(x, S) - \beta^\star_v(x, S) \;=\; \big(s^\star_u - s^\star_v\big) - \big(g_u(x) - g_v(x)\big), \qquad u, v \in S. \tag{4}$$

The quantity $\Delta_{uv}(x, S)$ is invariant to any shortlist-shared additive constant in $\beta^\star$ and captures exactly the correction difference needed to recover the Bayes ordering of the pair (u, v).

Definition 4.1 (Class-separable and non-class-separable regimes)**.** *The setting is* class-separable *if there exists a vector $a \in \mathbb{R}^K$ such that for every covered (x, S) and every $u, v \in S$,*

$$\Delta_{uv}(x, S) = a_u - a_v. \tag{5}$$

Equivalently, there exists a shortlist-shared scalar $c(x, S)$ such that $\beta^\star_y(x, S) = a_y + c(x, S)$ for all $y \in S$. If no such vector a exists, the setting is non-class-separable.

Intuitively, class-separable means that the ranking bias of each class can be summarized by a single number. Prior-based logit adjustment assumes exactly this structure: the correction depends only on the training frequency of the class, not on which rivals appear or how the input looks.

Proposition 4.2 (Classwise correction as an exact special case)**.** *If the setting is class-separable, then the fixed-offset score $r_y(x, S) = g_y(x) + a_y$ induces the Bayes-optimal ordering on every covered example.*

This positions logit adjustment precisely: it is exact when the residual correction is class-separable up to a shortlist-shared constant. Proof in Appendix D.4.3. The next subsection shows why this condition fails in general.

4.2 Why a pairwise correction is necessary

For classes $u, v \in [K]$, define the *base-score threshold*

$$t(x;\, u, v) \;:=\; g_v(x) - g_u(x). \tag{6}$$

A fixed classwise model ranks u above v if and only if $a_u - a_v > t(x;\, u, v)$. Thus a single cut $a_u - a_v$ must satisfy the requirements of *all* contexts in which the pair (u, v) appears.

Theorem 4.3 (Necessity of pairwise correction)**.** *Suppose there exist two covered contexts (x, S) and (x', S') and two labels $u, v \in S \cap S'$ such that the Bayes rule ranks u above v at x, ranks v above u at x', and*

$$t(x;\, u, v) \;\geq\; t(x';\, u, v). \tag{7}$$

Then no fixed offset vector $a \in \mathbb{R}^K$ can agree with the Bayes ordering in both contexts.

Theorem 4.3 identifies the precise point at which fixed classwise correction breaks down. Input-dependent pairwise variation alone does not force failure; the critical event is a *threshold crossing* for the same pair across contexts. We call such a pair *contradictory*. A setting may be non-class-separable without containing a contradictory pair, but once such a pair appears, no fixed classwise offset can agree with the Bayes ordering in all relevant contexts. In practice, a dataset can have large per-class score variance yet still be fully correctable by a fixed offset, as long as the variation does not reverse the ordering of any pair. Proof in Appendix D.4.4.

Threshold dispersion. Theorem 4.3 shows that threshold crossings drive the failure of fixed offsets. To quantify how prone a class is to such crossings, we define

$$D_y(S) \;:=\; \max_{j \in S \setminus \{y\}} \mathrm{std}\big(t(x;\, y, j)\big), \tag{8}$$

where the standard deviation is computed across all covered contexts in which both y and j appear in the shortlist. Large D_y means the true label faces at least one rival whose score gap fluctuates widely across inputs: some inputs are easy to distinguish, others are not, and no single fixed offset can handle both. In Section 6, we compute the mean threshold dispersion per rare class across all its covered test examples, then sort rare classes into five equal-sized quintiles (Q1–Q5, lowest to highest; we call these *contradiction quintiles* since high D_y is a necessary condition for threshold crossings) and show that the gain of REPAIR over classwise-only increases from Q1 to Q5.

4.3 Measuring reranking gains

Since all evaluation is conditioned on coverage, we normalize reranking gains by the recoverable gap within the shortlist:

$$\rho_k(\hat{y}) \;:=\; \frac{\mathrm{Hit@1}(\hat{y}) \;-\; \mathrm{Hit@1}(\hat{y}_{\mathrm{base}})}{1 \;-\; \mathrm{Hit@1}(\hat{y}_{\mathrm{base}})}, \tag{9}$$

where $\mathrm{Hit@1}(\hat{y}) := P(\hat{y}(X) = Y \mid C)$. The denominator is the fraction of covered examples the base model fails to rank first; $\rho_k = 1$ means the reranker corrects all such failures. We use ρ_k primarily in the synthetic validation below; for real-data experiments, we report Hit@1 and HFR directly since they are more interpretable across datasets with different coverage rates. The equivalent unconditional form is given in Appendix D.3.

4.4 Synthetic validation

To verify the predictions of Proposition 4.2 and Theorem 4.3, we construct two controlled settings: one where the residual is class-separable, and one where confuser pairs induce threshold crossings (setup in Appendix G.1). Figure 2 confirms both predictions: in the class-separable setting, classwise reranking and REPAIR achieve the same ρ_k ($= 0.17$); in the non-class-separable setting with contradictory pairs, REPAIR closes substantially more of the recoverable gap ($\rho_k = 0.42$ vs. 0.18).

5 Algorithm

Theorem 4.3 identifies the missing degree of freedom: the correction for a label cannot depend only on the label identity; it must also vary with the competing labels on the shortlist. More precisely, the excess risk of any plug-in reranker is controlled by how well it estimates the shortlist posterior (Appendix D.4.2), which in turn depends on approximating both the classwise and pairwise components of the residual. This section translates that structural insight into a concrete, lightweight algorithm.

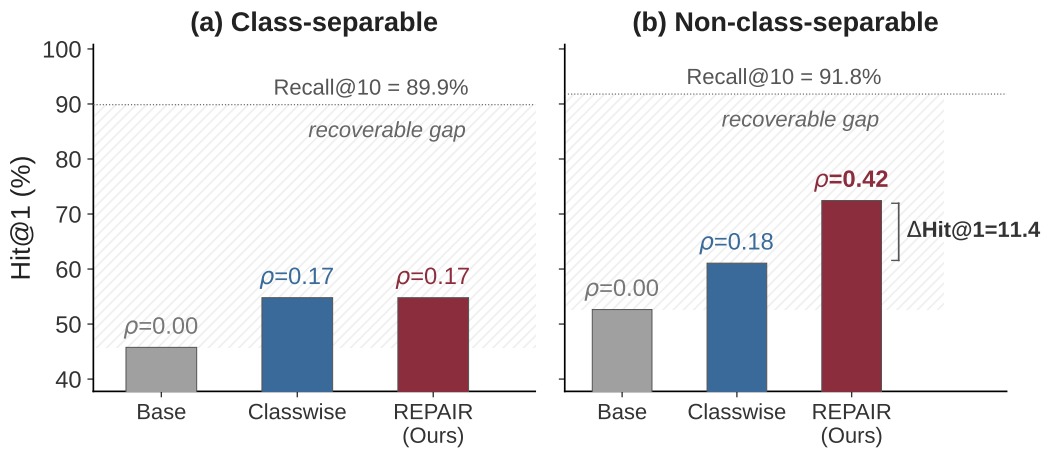


Figure 2: Synthetic validation (K=100, k=10). **(a)** Class-separable regime: classwise correction suffices ($\rho_k = 0.17$ for both), confirming Proposition 4.2. **(b)** Non-class-separable regime with contradictory pairs: REPAIR closes 2.3× the recoverable gap ($\rho_k = 0.42$ vs. 0.18), consistent with Theorem 4.3.

5.1 Score model

We model the reranker score as

$$r_y(x, S) \;=\; g_y(x) \;+\; \underbrace{a_y}_{\text{classwise}} \;+\; \underbrace{\ell_y(x, S)}_{\text{pairwise}}, \qquad y \in S, \tag{10}$$

where $\hat{\beta}_y := a_y + \ell_y(x, S)$ is our estimate of the residual correction $\beta^\star_y$ (Eq. 3). The classwise term a_y targets the class-separable component of $\beta^\star_y$ (Definition 4.1); the special choice $a_y = -\tau \log \pi_y$ recovers logit adjustment, but we learn a_y from covered calibration examples to obtain the strongest fixed-classwise correction. Even without pairwise term, this learned classwise correction outperforms the closed-form logit adjustment offset on most datasets (Table 1), because it captures per-class biases beyond frequency. The pairwise term ℓ_y targets the remainder that fixed classwise corrections structurally miss (Theorem 4.3).

The pairwise correction averages over all rivals in the shortlist:

$$\ell_y(x, S) \;=\; \frac{1}{|S| - 1} \sum_{j \in S \setminus \{y\}} \theta^\top \phi(x, y, j). \tag{11}$$

Each shortlisted rival j contributes a correction to label y through the shared parameter vector θ and the pairwise feature vector $\phi(x, y, j)$.

Two operational modes. This decomposition gives REPAIR two modes rather than a single fixed recipe. When the calibration data behave class-separably (Definition 4.1), the learned classwise term a_y alone recovers the Bayes ordering and the pairwise term can be omitted; this is the *Classwise mode*. When the residual is non-class-separable, the *Pairwise mode* adds $\ell_y(x, S)$ to capture the competition structure a fixed offset cannot. Classwise is therefore not merely a baseline but the class-separable special case of the same framework, and the pairwise term is optional: it is warranted precisely when a diagnostic check on calibration data shows a non-class-separable behavior.

5.2 Pairwise features

The feature vector $\phi(x, y, j)$ encodes how label y competes against rival j in context x. Four *competition features* are computed from the base model's output: the *score gap* $g_y(x) - g_j(x)$

(the negative of the threshold $t(x; y, j)$ in Eq. 6), the *rank gap* $\mathrm{rank}(j) - \mathrm{rank}(y)$, the *log-frequency ratio* $\log n_y - \log n_j$, where n_y denotes the number of training examples for class y, and the log-probability ratio $\log P(y|S)/P(j|S)$. *Label-similarity features* encode how confusable two classes are: taxonomic distance for iNaturalist, WordNet path similarity for ImageNet-LT, and HPO phenotype similarity for the rare-disease benchmarks (Appendix E.2).

5.3 Training and shrinkage

All parameters are estimated on *covered* examples from a held-out calibration set with $(y_i \in S_i)$. The base model is never modified. We fit a and θ jointly by maximizing a conditional multinomial logit on the shortlist:

$$q_{\theta,a}(y \mid x, S) \;=\; \frac{\exp\big(r_y(x,S)\big)}{\sum_{j\in S}\exp\big(r_j(x,S)\big)}, \quad y \in S. \tag{12}$$

The training objective is the regularized conditional log-likelihood:

$$\mathcal{L}(a,\theta;\mathcal{D}) \;=\; \sum_{(x_i,S_i,y_i)\in\mathcal{D}} \log q_{\theta,a}(y_i \mid x_i, S_i) \;-\; \lambda_\theta\|\theta\|_2^2 \;-\; \lambda_a\|a\|_2^2, \tag{13}$$

where $\mathcal{D} = \{(x_i, S_i, y_i) : y_i \in S_i\}$ is the set of covered examples.

Shrinkage. For rare classes with very few covered examples, a_y can be noisy. We stabilize via empirical-Bayes shrinkage toward a frequency-group mean:

$$a_y^\star \leftarrow (1-\lambda_y)\,\hat{a}_y + \lambda_y\,\mu_{b(y)}, \qquad \lambda_y = \frac{\sigma_y^2}{\sigma_y^2 + \nu_{b(y)}^2}, \tag{14}$$

where $b(y)$ indexes a frequency group, $\mu_{b(y)}$ is the group mean, and σ_y^2, $\nu_{b(y)}^2$ are within-class and between-group variance estimates (Appendix D.4.7). The complete procedure is summarized in Algorithm 1 (Appendix E.1).

6 Experiments

6.1 Setup

Datasets and base models. We evaluate on nine benchmarks spanning text classification, image classification, species recognition, and rare-disease diagnosis. The four pure-text NLP benchmarks are GOEMOTIONS (28 classes), MASSIVE (60 classes), HUFFPOST-LT (42 classes), and FEW-NERD (66 classes), each using a Qwen3.5-2B base classifier fine-tuned with LoRA. The remaining benchmarks are image classification (IMAGENET-LT, 1,000 classes; PLACES-LT, 365 classes), species recognition (INATURALIST 2018, 8,142 classes), and rare-disease diagnosis (GMDB, 508 classes; RAREBENCH, 508 classes, out-of-distribution). Base models range from ResNet-50/152 with cRT classifiers to a fine-tuned Qwen3.5-0.8B for the clinical benchmarks and Qwen3.5-2B for the NLP benchmarks; full details are in Appendix F.1. In all cases, REPAIR operates on a fixed top-k shortlist (k=10) produced by the base model; the base model is never modified. We designate the bottom 80% of classes by training-set frequency as *rare* and the top 20% as *frequent*.

Baselines. All methods share the same shortlist. **Base** ranks by the raw scores $g_y(x)$. **LogitAdj** (Menon et al., 2021) adds $-\tau \log \pi_y$ with τ tuned on the calibration set. **τ-norm** (Kang et al., 2020) divides by the classifier weight norm $\|w_y\|^\tau$. **Classwise** learns per-class offsets a_y via Eq. 13 with shrinkage (Eq. 14), setting $\theta = 0$. **REPAIR** adds the pairwise term $\ell_y(x, S)$.

Evaluation schemes. Reranker parameters are estimated on a held-out calibration set and evaluated on an independent test set; the two never overlap. Each experiment is repeated

over five trials, each subsampling 80% of the calibration set with a distinct seed. We report mean $\pm$ std for methods with learned parameters (Classwise, REPAIR) and single values for LogitAdj and τ-norm, whose corrections are deterministic given the training-set frequency distribution and a fixed τ (tuned once on the full calibration set), and therefore do not vary across calibration subsamples. Significance is assessed via a sign test ($\geq 4/5$ wins); full results are in Appendix G.7. All metrics are conditioned on coverage ($Y \in S$). We report Hit@1, Hit@3, MRR, Rare Hit@1, Frequent Hit@1, and the hardest-rival flip rate (HFR; Definition 6.1), which measures whether the reranker can beat the strongest wrong competitor on inputs the base model misranks.

Definition 6.1 (Hardest-rival flip rate)**.** *The probability that the reranker ranks the true label above the base model's top-scoring wrong class, among covered examples the base model misranks:*

$$\mathrm{HFR}(\hat{y}) = P\big(r_Y > r_{h(X,Y)} \mid C,\ \hat{y}_{\text{base}} \neq Y\big),$$

where $h(x, y) := \arg\max_{j \in S(x) \setminus \{y\}} g_j(x)$.

6.2 Results

Table 1 reports the comparison on the vision and clinical benchmarks, and Table 10 the pure-text NLP benchmarks. The decomposition makes a direct prediction: on datasets where the base model's score gaps are stable across inputs, classwise correction should suffice and the pairwise term should add little; on datasets where score gaps vary widely, the pairwise term should provide substantial gains.

Near-class-separable regime (vision benchmarks). On iNaturalist, ImageNet-LT, and Places-LT, REPAIR provides small but consistent Hit@1 improvements over Classwise (+0.2%, +0.2%, +0.4%), with the learned Classwise correction competitive with or outperforming LogitAdj on all three. HFR nearly doubles on iNaturalist (0.074 $\rightarrow$ 0.138) and ImageNet-LT (0.336 $\rightarrow$ 0.378). On Places-LT, LogitAdj and τ-norm both degrade Hit@1 relative to Base, while Classwise and REPAIR both exceed it, confirming that learned offsets are more robust than closed-form corrections. The decomposition identifies these datasets as predominantly classwise: most ranking errors are correctable by a fixed per-class offset.

Non-class-separable regime (rare-disease benchmarks). On GMDB, REPAIR improves Hit@1 from 71.4 to 72.3 and HFR from 0.441 to 0.510. The aggregate Hit@1 gain is driven by the model-derived competition features rather than by HPO specifically; removing the HPO feature leaves every REPAIR metric unchanged by less than 0.1 (Appendix G.10). HPO instead shapes the composition of the top-k differential diagnosis, promoting the correct diagnosis while keeping phenotypically related conditions near the top (Appendix G.11). On RareBench (OOD), Hit@1 rises from 66.2 to 85.6 (+19.4 %) and HFR from 0.215 to 0.708. Two factors drive this: the base model operates out-of-distribution (trained on GMDB with image and text, tested on text), which amplifies threshold variation across inputs; and HPO phenotype similarity gives the pairwise term discriminative signal the base model has no access to. The RareBench test set contains only 32 covered examples, so we verify this result with 10K-resample bootstrap: the 95% CI for the gain over Classwise is $[+6.2,\ +31.2]$(%)(P=0.999; Appendix G.8).

Pure-text NLP classifiers. The decomposition is modality-agnostic: it applies to any classifier that produces class scores, including language models with a classification head. We verify this on four pure-text NLP benchmarks (GoEmotions, MASSIVE, HuffPost-LT, Few-NERD), each using a Qwen3.5-2B base classifier fine-tuned with LoRA under the identical protocol. The same contrast holds: on the class-separable GoEmotions, Classwise already suffices and REPAIR adds little (+0.05 Hit@1 over Classwise), whereas on the non-class-separable MASSIVE, HuffPost-LT, and Few-NERD the pairwise term yields clear gains (+1.78, +0.84, +1.05 Hit@1, and +5.53, +2.04, +1.83 Rare Hit@1 over Classwise). Full results are in Appendix G.9.

Head–tail trade-off. On vision benchmarks, improving Rare Hit@1 comes at some cost to Frequent Hit@1. Most of this shift originates from the classwise correction itself; the pairwise term adds only a small further change. On RareBench, both strata improve simultaneously.

Component ablation. In the synthetic setting (Figure 6, Appendix G.3), CW-only matches REPAIR in the class-separable regime while PW-only adds nothing; in the non-class-separable regime, the full model outperforms both (72.5 vs. 61.1 CW-only, 72.1 PW-only). On real data (Appendix G.3), the relative importance of the two components shifts across the spectrum. On iNaturalist, PW-only alone matches REPAIR (Rare Hit@1: 42.1), suggesting that the pairwise features implicitly absorb the classwise correction. On ImageNet-LT, Places-LT and GMDB, both components contribute incrementally. On RareBench, neither component alone approaches the full model (CW-only 58.8, PW-only 47.1, REPAIR 88.2; single random seed): the classwise term calibrates the score scale while the pairwise term redirects rankings, and joint estimation is essential for the two corrections to reinforce each other.

6.3 Quintile analysis

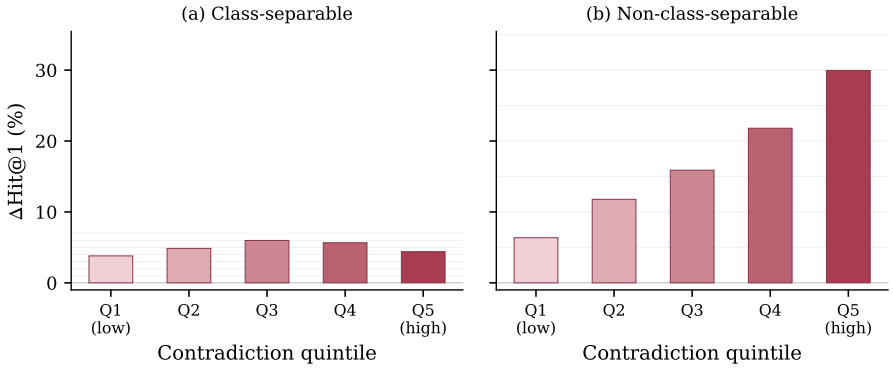


Figure 3: ΔHit@1 of REPAIR over Classwise by mean D_y quintile (synthetic, K=100, k=10, rare classes). **(a)** Class-separable: almost flat across quintiles. **(b)** Non-class-separable: monotone increasing from Q1 to Q5.

We stratify rare classes by their mean threshold dispersion D_y (Eq. 8) into five equal-sized quintiles and measure the ΔHit@1 of REPAIR over Classwise within each bin. In the class-separable setting (a), the gain is uniformly small and flat, which means D_y does not modulate the pairwise contribution when the residual is approximately classwise. In the non-class-separable setting (b), the gain increases monotonically from Q1 (+6%) to Q5 (+30%). On real data (Appendix G.4), the same pattern holds: on GMDB, Q5 shows the largest gain while Q1 shows the smallest, consistent with the non-class-separable regime; on iNaturalist, the gain is uniformly small and the quintile gradient is flat, consistent with a near-class-separable residual. ImageNet-LT and Places-LT fall between these extremes.

7 Conclusion

In this work, we proposed a residual decomposition framework for post-hoc reranking under long-tailed class distributions. Unlike logit adjustment, which applies a fixed per-class offset derived from class frequencies, our framework decomposes the Bayes-optimal correction into a learned classwise term and an input-dependent pairwise term, with a formal characterization of when the classwise component alone is sufficient and when it is not. Empirically, the learned classwise correction already outperforms logit adjustment on most datasets, and the pairwise term provides further gains proportional to the degree of threshold variation across nine benchmarks spanning text classification, visual recognition, and rare-disease diagnosis, with especially large improvements in the rare-disease setting where phenotypically similar conditions create strong pairwise competition. The current

Table 1: Reranking results on three visual recognition and two rare-disease diagnosis benchmarks (shortlist size $k = 10$). All metrics are conditioned on the true label appearing in the shortlist; arrows indicate higher is better. Rare and Frequent denote the bottom 80% and top 20% of classes by training-set frequency. Classwise is our framework without the pairwise term; both Classwise and REPAIR report mean$\pm$std over 5 calibration subsamples. Best per dataset in **bold**. Row shading: blue = near-class-separable (vision benchmarks), red = non-class-separable (rare-disease benchmarks).

			Overall			*Stratified Hit@1*		
Task	Dataset	Method	Hit@1↑	Hit@3↑	MRR↑	Rare↑	Freq.↑	HFR↑
Species Recog.	iNATURALIST (ResNet-50)	Base	47.0	71.7	.627	37.5	**60.0**	–
		LogitAdj	**49.0**	74.7	.648	41.7	58.9	.097
		τ-norm	48.1	73.7	.640	40.0	59.2	.105
		Classwise	48.8$\pm$.02	74.5$\pm$.01	.646$\pm$.000	41.1$\pm$.02	59.3$\pm$.01	.074$\pm$.001
		REPAIR (Ours)	**49.0**$\pm$.01	**75.6**$\pm$.02	**.652**$\pm$.000	**42.2**$\pm$.02	58.4$\pm$.02	**.138**$\pm$.002
Image Classif.	IMAGENET-LT (ResNet-50)	Base	54.4	78.6	.689	47.4	**76.7**	–
		LogitAdj	61.5	85.3	.748	59.5	68.1	.321
		τ-norm	60.5	84.8	.741	58.1	68.2	.299
		Classwise	61.8$\pm$.02	85.4$\pm$.02	.750$\pm$.000	59.9$\pm$.05	67.8$\pm$.1	.336$\pm$.001
		REPAIR (Ours)	**62.0**$\pm$.04	**85.4**$\pm$.03	**.751**$\pm$.000	**60.8**$\pm$.05	65.7$\pm$.1	**.378**$\pm$.001
Scene Recog.	PLACES-LT (ResNet-152)	Base	44.7	72.1	.617	46.3	**37.9**	–
		LogitAdj	44.5	71.5	.613	47.9	30.4	.072
		τ-norm	19.7	45.5	.397	24.5	0.0	**.405**
		Classwise	45.1$\pm$.1	72.4$\pm$.2	.620$\pm$.000	46.9$\pm$.1	37.4$\pm$.3	.028$\pm$.001
		REPAIR (Ours)	**45.5**$\pm$.1	**72.5**$\pm$.1	**.622**$\pm$.001	**48.2**$\pm$.1	34.1$\pm$.3	.072$\pm$.001
Rare Disease Diagnosis	GMDB (Qwen-0.8B)	Base	64.0	79.3	.741	62.9	64.7	–
		LogitAdj	70.6	85.5	.797	78.8	65.1	.422
		τ-norm	67.4	81.6	.769	66.4	**68.0**	.168
		Classwise	71.4$\pm$.1	85.8$\pm$.0	.801$\pm$.001	78.5$\pm$.1	66.5$\pm$.2	.441$\pm$.004
		REPAIR (Ours)	**72.3**$\pm$.3	**85.8**$\pm$.2	**.806**$\pm$.001	**79.6**$\pm$.3	67.4$\pm$.5	**.510**$\pm$.005
	RAREBENCH (Qwen-0.8B, text)	Base	59.4	75.0	.712	41.2	80.0	–
		LogitAdj	62.5	71.9	.712	52.9	73.3	.385
		τ-norm	56.2	78.1	.702	41.2	73.3	.000
		Classwise	66.2$\pm$1.2	91.2$\pm$2.3	.789$\pm$.006	54.1$\pm$2.4	80.0$\pm$.0	.215$\pm$.031
		REPAIR (Ours)	**85.6**$\pm$4.2	**92.5**$\pm$1.5	**.899**$\pm$.026	**81.2**$\pm$4.4	**90.7**$\pm$5.3	**.708**$\pm$.090

pairwise features are compact and domain-interpretable, but richer representations learned from the base model's embeddings could capture competition patterns that hand-crafted features miss. Overall, these results demonstrate that in long-tailed settings, decomposing the residual correction into classwise and pairwise components offers a principled and lightweight alternative to purely frequency-based post-hoc adjustments, particularly in domains where local competition between similar classes causes ranking errors that no fixed offset can resolve.

Acknowledgments

This study is supported by NIH grant *OD037960* and the Children's Hospital of Philadelphia (CHOP) Research Institute. We also gratefully acknowledge the Penn Advanced Research Computing Center (PARCC) for providing computational resources and support.

Ethics Statement

This work applies post-hoc reranking to rare-disease diagnosis, a high-stakes clinical setting. REPAIR is designed as a decision-support tool that reorders a candidate shortlist; it does not make final diagnostic decisions. This study is approved by the institutional review board at the Children's Hospital of Philadelphia (CHOP) under protocol *IRB 18-015712*. All datasets used in this work are publicly available.

References

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019.

Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen. Rarebench: Can llms serve as rare diseases specialists? In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9260–9269, 2019.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054, 2020.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. Few-NERD: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3198–3213, 2021.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4277–4302, 2023.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Yaron Gurovich, Yair Hanani, Omri Bar, Guy Nadav, Nicole Fleischer, Dekel Gelbman, Lina Basel-Salmon, Peter M. Krawitz, Susanne B. Kamphausen, Martin Zenker, et al. Identifying facial phenotypes of genetic disorders using deep learning. *Nature Medicine*, 25:60–64, 2019.

Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6622–6632, 2021.

Tzung-Chien Hsieh, Aviram Bar-Haim, Shahida Moosa, Nadja Ehmke, Karen W. Gripp, Jean Tori Pantel, Magdalena Danyel, Martin Atta Mensah, Denise Horn, Stanislav Rosnev, et al. Gestaltmatcher facilitates rare disease matching using facial phenotype descriptors. *Nature Genetics*, 54:349–357, 2022.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

Byungju Kim and Junmo Kim. Adjusting decision boundary for class imbalanced learning. *IEEE Access*, 8:81674–81685, 2020.

Hellen Lesmann, Alexander Hustinx, Shahida Moosa, Hannah Klinkhammer, Elaine Marchi, Pilar Caro, Ibrahim M. Abdelrazek, Jean Tori Pantel, Merle ten Hagen, Meow-Keong Thong, et al. Gestaltmatcher database - a global reference for facial phenotypic variability in rare human diseases. *medRxiv preprint*, 2024. doi: 10.1101/2023.06.06.23290887.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2): 318–327, 2020.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2532–2541, 2019.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

Rishabh Misra. News category dataset. *arXiv preprint arXiv:2209.11429*, 2022.

Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems*, 2020.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021.

Zhanliang Wang, Da Wu, Quan Nguyen, and Kai Wang. Integrating chain-of-thought and retrieval augmented generation enhances rare disease diagnosis from clinical notes. *Medicine Bulletin*, 2(2):167–183, 2026a.

Zhanliang Wang, Da Wu, Quan M. Nguyen, Zhuoran Xu, and Kai Wang. Multimodal integrated knowledge transfer to large language models through preference optimization with biomedical applications. *Patterns*, 2026b. ISSN 2666-3899. doi: 10.1016/j.patter.2026.101544. URL https://doi.org/10.1016/j.patter.2026.101544. Article 101544.

Da Wu, Zhanliang Wang, Hongzhuo Chen, Jingye Yang, Cong Liu, Tzung-Chien Hsieh, Elaine Marchi, Justin Blair, Peter Krawitz, Chunhua Weng, Wendy Chung, Gholson J. Lyon, Ian D. Krantz, Jennifer M. Kalish, and Kai Wang. Gestaltmml: Enhancing rare genetic disease diagnosis through multimodal machine learning combining facial images and clinical text, 2026. URL https://arxiv.org/abs/2312.15320.

Jiancong Xiao, Bojian Hou, Zhanliang Wang, Ruochen Jin, Qi Long, Weijie J Su, and Li Shen. Restoring calibration for aligned large language models: A calibration-aware fine-tuning approach. In *Proceedings of 42th International Conference on Machine Learning*, 2025.

Lu Yang, He Jiang, Qing Song, and Jun Guo. A survey on long-tailed visual recognition. *International Journal of Computer Vision*, 130(7):1837–1872, 2022.

Yifan Zhang, Bryan Hooi, Lanqing HONG, and Jiashi Feng. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. In *Advances in Neural Information Processing Systems*, 2022.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep Long-Tailed Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10795–10816, 2023.

Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16484–16493, 2021.

A Use of Generative AI

We used large language models (Claude Opus 4.6) to polish grammar and improve the clarity of our writing, to organize the structure and logical flow of the paper, and to check and refine the presentation of mathematical proofs. All scientific contributions, experimental design, and final content were produced and verified by the authors.

B Additional Related Work

Training-time corrections. Training time methods modify the loss, sampling or architecture during learning. Re-weighing approaches adjust per-class loss contributions based on class frequency or sample difficulties (Lin et al., 2020; Cui et al., 2019), while margin-based methods enforce larger-decision margin for rare classes (Cao et al., 2019). At sampling level, decoupled training learns representation with standard sampling and re-balances the classifier (Kang et al., 2020; Zhong et al., 2021). Ensemble methods in addition offer structural solution by training multiple experts with complementary skills across frequent classes (Zhang et al., 2022; Wang et al., 2021).

C Notation Reference

Table 2 summarizes the principal symbols used throughout the paper.

D Theoretical Supplement

D.1 Restricted posterior on the shortlist

Because $S(x)$ is a deterministic function of x, the coverage probability given $X = x$ is

$$\alpha(x) := P(C \mid X = x) = \sum_{j \in S(x)} P(Y = j \mid X = x). \tag{15}$$

Whenever $\alpha(x) > 0$,

$$p_S(y \mid x) = P(Y = y \mid X = x, C) = \frac{P(Y = y \mid X = x) \cdot \mathbf{1}\{y \in S(x)\}}{\alpha(x)}. \tag{16}$$

Hence the Bayes-optimal shortlist score is

$$s^\star_y(x, S) = \log p_S(y \mid x) = \log P(Y = y \mid X = x) - \log \alpha(x), \qquad y \in S(x), \tag{17}$$

so only the relative values across $y \in S(x)$ matter for ranking. The term $\log \alpha(x)$ is shared by all labels in the shortlist and therefore cancels out of any pairwise comparison.

D.2 Weighted variant

The paper uses ordinary top-1 risk for clarity. If one instead optimizes a class-weighted risk with weights $w_y > 0$, then the Bayes-optimal shortlist score becomes $\log w_y + \log p_S(y \mid x)$. All main arguments carry over verbatim after adding $\log w_y$ to the ideal score. We keep the unweighted version in the main text because it is easier to read.

Table 2: Summary of notation.

Symbol	Definition	Location
Setup		
X	Input	§3
$Y \in [K]$	True label	§3
π_y	Class prior $P(Y{=}y)$	§3
$g_y(x)$	Base-model score for class y	§3
k	Shortlist size	§3
$S(x)$	Top-k shortlist under g	§3
C	Coverage event $\{Y \in S(X)\}$	§3
Theory (§4)		
$p_S(y \mid x)$	Shortlist posterior	Eq. (1)
$s^\star_y(x, S)$	Bayes-optimal shortlist score	Eq. (2)
$\beta^\star_y(x, S)$	Residual correction $s^\star_y - g_y$	Eq. (3)
$\Delta_{uv}(x, S)$	Pairwise gap $\beta^\star_u - \beta^\star_v$	Eq. (4)
$t(x;\, u, v)$	Base-score threshold $g_v(x) - g_u(x)$	Eq. (6)
$D_y(S)$	Threshold dispersion	Eq. (8)
ρ_k	Recoverable gap closure	Eq. (9)
Algorithm (§5)		
$r_y(x, S)$	Reranker score $g_y + \hat{\beta}_y$	Eq. (10)
$\hat{\beta}_y$	Estimated residual $a_y + \ell_y$	§5.1
a_y	Learned classwise offset	Eq. (10)
$\ell_y(x, S)$	Pairwise correction	Eq. (10)
θ	Shared pairwise parameter	Eq. (11)
$\phi(x, y, j)$	Pairwise feature vector	§5.2
$\nu^2_{b(y)}$, σ^2_y	Shrinkage variances	Eq. (14)
Evaluation (§6)		
Hit@k	$P(\mathrm{rank}_r(Y) \le k \mid C)$	§6.1
MRR	$\mathbb{E}[\mathrm{rank}_r(Y)^{-1} \mid C]$	§6.1
HFR	Hardest-rival flip rate	Def. 6.1

D.3 Unconditional form of the recoverable gap closure

When expressed in unconditional accuracy, the recoverable gap closure (9) takes the equivalent form

$$\rho_k(\hat{y}) = \frac{\mathrm{Acc}(\hat{y}) - \mathrm{Acc}(\hat{y}_{\mathrm{base}})}{P(C) - \mathrm{Acc}(\hat{y}_{\mathrm{base}})}, \tag{18}$$

since $\mathrm{Hit@1}(\hat{y}) = \mathrm{Acc}(\hat{y})/P(C)$ and the $P(C)$ factors cancel. This equivalence holds because all rerankers share the same shortlist and hence the same coverage $P(C)$.

D.4 Proofs

D.4.1 Bayes-optimal reranking on the shortlist

Proposition D.1 (Bayes rule on a fixed shortlist)**.** *Among all measurable rules $\hat{y}(X) \in S(X)$, a Bayes-optimal reranker can be chosen pointwise as $\hat{y}^\star(x) \in \arg\max_{y \in S(x)} p_S(y \mid x)$. Equivalently, any score function strictly increasing in $p_S(y \mid x)$ yields the same Bayes ordering.*

Proof. Fix x. Among all measurable rules $r(x) \in S(x)$, the conditional probability of being correct is

$$P(r(X) = Y \mid X = x, C) = \sum_{y \in S(x)} \mathbf{1}\{r(x) = y\}\, p_S(y \mid x) = p_S(r(x) \mid x).$$

Since $r(x)$ is a deterministic choice, the indicator is nonzero for exactly one value of y, and the sum collapses to $p_S(r(x) \mid x)$. Therefore the conditional risk given $X = x$ and C is minimized by any $r(x) \in \arg\max_{y \in S(x)} p_S(y \mid x)$. Since this argument is pointwise in x, any measurable selector of that argmax is Bayes-optimal on covered examples. □

D.4.2 Plug-in estimation error

Proposition D.2 (Plug-in control)**.** *Let $\hat{p}_S(\cdot \mid x)$ be an estimate of $p_S(\cdot \mid x)$, and let $\hat{y}_{\text{plug}}(x) \in \arg\max_{y \in S(x)} \hat{p}_S(y \mid x)$. Then the conditional excess risk on covered examples satisfies*

$$P(\hat{y}_{\text{plug}}(X) \neq Y \mid C) - P(\hat{y}^\star(X) \neq Y \mid C) \leq 2\,\mathbb{E}[\|\hat{p}_S(\cdot \mid x) - p_S(\cdot \mid x)\|_1 \mid C]. \tag{19}$$

Lemma D.3 (Covered excess-risk identity)**.** *For any measurable shortlist rule $\hat{y}(X) \in S(X)$,*

$$P(\hat{y}(X) \neq Y \mid C) - P(\hat{y}^\star(X) \neq Y \mid C) = \mathbb{E}[p_S(\hat{y}^\star(X) \mid x) - p_S(\hat{y}(X) \mid x) \mid C].$$

Proof. Condition on $X = x$ and C. The conditional error rates are $1 - p_S(\hat{y}(x) \mid x)$ and $1 - p_S(\hat{y}^\star(x) \mid x)$, respectively. Subtracting:

$$\big(1 - p_S(\hat{y}(x) \mid x)\big) - \big(1 - p_S(\hat{y}^\star(x) \mid x)\big) = p_S(\hat{y}^\star(x) \mid x) - p_S(\hat{y}(x) \mid x).$$

Taking expectations over X given C yields the identity. □

Proof of Proposition (Plug-in control). Using the lemma,

$$\Delta := P(\hat{y}_{\text{plug}}(X) \neq Y \mid C) - P(\hat{y}^\star(X) \neq Y \mid C) = \mathbb{E}\big[p_S(\hat{y}^\star(X) \mid x) - p_S(\hat{y}_{\text{plug}}(X) \mid x) \mid C\big].$$

For each x, we add and subtract $\hat{p}_S$:

$$\begin{aligned} p_S(\hat{y}^\star(x) \mid x) - p_S(\hat{y}_{\text{plug}}(x) \mid x) = {} & \underbrace{p_S(\hat{y}^\star(x) \mid x) - \hat{p}_S(\hat{y}^\star(x) \mid x)}_{\text{(i)}} \\ & + \underbrace{\hat{p}_S(\hat{y}^\star(x) \mid x) - \hat{p}_S(\hat{y}_{\text{plug}}(x) \mid x)}_{\text{(ii)}} \\ & + \underbrace{\hat{p}_S(\hat{y}_{\text{plug}}(x) \mid x) - p_S(\hat{y}_{\text{plug}}(x) \mid x)}_{\text{(iii)}}. \end{aligned}$$

Term (ii) ≤ 0 by definition of $\hat{y}_{\text{plug}}$. Each of terms (i) and (iii) is bounded by $\|\hat{p}_S(\cdot \mid x) - p_S(\cdot \mid x)\|_1$. Taking conditional expectations given C completes the proof. □

D.4.3 Proof of Proposition 4.2

Proof. If $\beta^\star_y(x, S) = a_y + c(x, S)$ for every covered triple (x, S, y) and some shortlist-shared scalar $c(x, S)$, then

$$s^\star_y(x, S) = g_y(x) + a_y + c(x, S) \quad \text{for all } y \in S.$$

Therefore the fixed-offset score $g_y(x) + a_y$ differs from the Bayes-optimal shortlist score only by the shortlist-shared constant $c(x, S)$. Since adding the same constant to all shortlisted labels does not change the ranking, the ordering induced by $g_y(x) + a_y$ is Bayes-optimal on every covered example. □

D.4.4 Proof of Theorem 4.3

Proof. Assume for contradiction that some fixed-classwise model $g_\cdot(x) + a_\cdot$ agrees with the Bayes ordering in both contexts.

Because the Bayes rule ranks u above v at x, the fixed-classwise model must satisfy $g_u(x) + a_u > g_v(x) + a_v$, equivalently

$$a_u - a_v > g_v(x) - g_u(x) = t(x;\, u, v). \tag{20}$$

Because the Bayes rule ranks v above u at x', the same model must satisfy $g_v(x') + a_v > g_u(x') + a_u$, equivalently

$$a_u - a_v < g_v(x') - g_u(x') = t(x';\, u, v). \tag{21}$$

Combining yields $t(x;\, u, v) < a_u - a_v < t(x';\, u, v)$, which requires $t(x;\, u, v) < t(x';\, u, v)$. But by assumption $t(x;\, u, v) \geq t(x';\, u, v)$, a contradiction. □

D.4.5 Attainable frontier under fixed retrieval

Remark 1. *For any reranker $\hat{y}(X) \in S(X)$, the event $\{\hat{y}(X) = Y\}$ implies C, so $P(\hat{y}(X) = Y) = P(\hat{y}(X) = Y, C) \leq P(C)$. Equivalently,* Hit@1$(\hat{y}) \leq 1$.

D.4.6 Conditional likelihood for the training objective

For covered training examples (x_i, S_i, y_i), the model

$$q_{\theta,a}(y \mid x, S) = \frac{\exp\big(r_y(x, S)\big)}{\sum_{j \in S} \exp\big(r_j(x, S)\big)}$$

is a standard multinomial logit model over the shortlist. Assuming conditional independence across covered examples, the log-likelihood is $\sum_{i:\, y_i \in S_i} \log q_{\theta,a}(y_i \mid x_i, S_i)$, and adding ℓ_2 penalties yields Eq. (13).

D.4.7 Empirical-Bayes shrinkage

A simple way to stabilize rare-class offsets is to treat the raw estimate $\tilde{a}_y$ as a noisy observation of an unobserved true offset a_y, and to place a Gaussian prior on a_y within a frequency group $b(y)$:

$$\tilde{a}_y \mid a_y \sim \mathcal{N}(a_y,\, \sigma_y^2), \qquad a_y \sim \mathcal{N}(\mu_{b(y)},\, \nu_{b(y)}^2).$$

The posterior mean is

$$\mathbb{E}[a_y \mid \tilde{a}_y] = (1 - \lambda_y)\, \tilde{a}_y + \lambda_y\, \mu_{b(y)}, \qquad \lambda_y = \frac{\sigma_y^2}{\sigma_y^2 + \nu_{b(y)}^2}. \tag{22}$$

The more uncertain the class-specific estimate $\tilde{a}_y$ is (large σ_y^2), the more it is pulled toward the group mean $\mu_{b(y)}$. For classes with many covered examples, σ_y^2 is small and the estimate stays close to the raw value; for rare classes with few examples, the estimate is regularized toward the group prior.

E Method Details

E.1 Algorithm pseudocode

E.2 Pairwise feature definitions

Table 3 gives the per-dataset pairwise feature specifications.

The score gap directly encodes the base-score threshold $t(x;\, y, j) = g_j(x) - g_y(x)$ from Theorem 4.3, allowing the pairwise correction to adapt to the current margin. The domain

Algorithm 1 REPAIR: Reranking via Pairwise Residual Correction

Require: Base scorer g, validation set $\{(x_i, y_i)\}_{i=1}^n$, shortlist size k
Training:
for $i = 1, \ldots, n$ **do**
  $S_i \leftarrow$ top-k classes by $g(x_i)$
end for
$\mathcal{D} \leftarrow \{(x_i, S_i, y_i) : y_i \in S_i\}$
$(a^*, \theta^*) \leftarrow \arg\max_{a,\theta} \mathcal{L}(a, \theta; \mathcal{D})$ ▷ joint optimization
for each class y **do**
  $a_y^* \leftarrow (1 - \lambda_y)\, a_y^* + \lambda_y\, \mu_{b(y)}$ ▷ shrinkage
end for
Inference:
Require: Input x, learned (a^*, θ^*)
$S \leftarrow$ top-k classes by $g(x)$
for $y \in S$ **do**
  $\ell_y \leftarrow \frac{1}{k-1} \sum_{j \in S \setminus \{y\}} {\theta^*}^\top \phi(x, y, j)$ ▷ pairwise correction
  $r_y \leftarrow \underbrace{g_y(x)}_{\text{base}} + \underbrace{a_y^*}_{\text{classwise}} + \underbrace{\ell_y}_{\text{pairwise}}$
end for
return $\arg\max_{y \in S} r_y$

Table 3: Pairwise feature vector $\phi(x, y, j)$ per dataset. All datasets share the first three competition features; the domain-specific similarity and log-frequency ratio vary.

Feature	iNat	ImageNet-LT	Places-LT	GMDB	RareBench	Dim
Score gap $g_y(x) - g_j(x)$	✓	✓	✓	✓	✓	1
Rank gap	✓	✓	✓	✓	✓	1
Log-prob ratio $\log \frac{P(y\|S)}{P(j\|S)}$	✓	✓	✓	✓	✓	1
Domain similarity	Taxonomic	WordNet	—	HPO Jaccard	HPO Jaccard	0–1
Log-freq ratio $\log \frac{n_y+1}{n_j+1}$	✓	✓	✓	✓	✓	1
Total dimension	5	5	4	5	5	

similarity feature captures how confusable two classes are independent of the current input: taxonomic distance in the biological hierarchy for iNaturalist, WordNet path similarity for ImageNet-LT, and HPO phenotype Jaccard similarity for the rare-disease benchmarks. Places-LT uses four features (no domain similarity available for scene categories).

E.3 Implementation details

Optimization. All reranker parameters are estimated using L-BFGS-B (`scipy.optimize.minimize`, maxiter=200–300, ftol=10^{-11}). Wall-clock time is under 30 seconds per dataset per seed on a single CPU.

Hyperparameters. Table 4 reports the per-dataset settings. All values were selected by grid search on the calibration set (maximizing Hit@1 among covered examples); the test set was never used for selection.

Hyperparameter search grid. $\tau \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$; $\lambda_a \in \{0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$; $\lambda_\theta \in \{5\times10^{-5}, 10^{-4}, 5\times10^{-4}, 10^{-3}, 5\times10^{-3}, 10^{-2}\}$; $\alpha \in \{0.0, 0.1, 0.3, 0.5, 0.7, 0.9\}$.

Table 4: Hyperparameter settings per dataset.

	iNat	ImageNet-LT	Places-LT	GMDB	RareBench
λ_a	0.01	0.001	0.001	0.5	0.005
λ_θ	0.005	0.001	0.001	0.01	5×10^{-5}
α	0.3	0.0	0.3	0.3	0.0
τ	1.0	1.0	0.0	0.0	0.0
Features	5	5	4	5	5
Shrinkage groups	1	1	1	1	1

Base models. For iNaturalist, we fine-tune a ResNet-50 (torchvision `IMAGENET1K_V2`) on the iNat 2018 training split. For ImageNet-LT, we use the publicly available ERM ResNet-50 checkpoint from Kang et al. (2020).[1] For Places-LT, we extract frozen ResNet-152 features (ImageNet-pretrained) and train a logistic-regression classifier with class-balanced weighting (cRT). For GMDB,[2] we select the 508 diseases (out of 528) that have facial phenotype HPO annotations and fine-tune Qwen3.5-0.8B,[3] a multimodal vision-language model, as a 508-way classifier on 7,844 facial images; base logits are pre-normalized by classifier weight norms ($g_y \leftarrow g_y / \|w_y\|^{0.5}$) before reranking. For RareBench,[4] the same GMDB-trained checkpoint is evaluated in a text-only setting: HPO phenotype terms are formatted as a clinical prompt with a blank 224×224 image placeholder, bypassing the visual modality entirely.

F Datasets and Evaluation

F.1 Dataset details

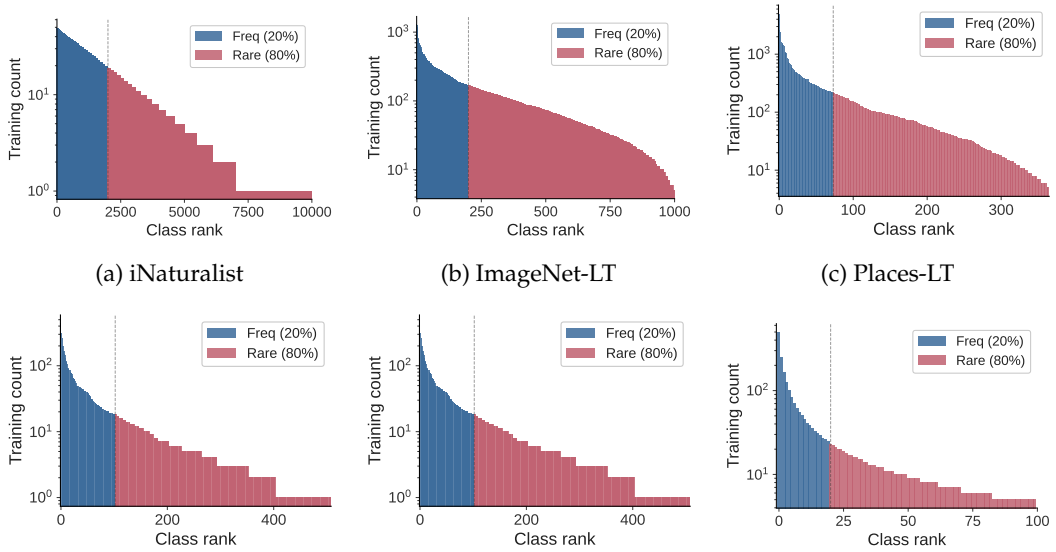
Table 5: Dataset overview. K = number of classes; Cal/Test = calibration and test set sizes; Rare/Freq = class counts under the uniform 80/20 split by training-set frequency. All rerankers operate on a k=10 shortlist.

Dataset	Base model	K	Cal	Test	Rare	Freq	Recall@10	Split source
iNaturalist	ResNet-50	8,142	30,000	70,000	6,514	1,628	37.8%	Original split
ImageNet-LT	ResNet-50	1,000	20,000	30,000	800	200	76.7%	Random (seed=42)
Places-LT	ResNet-152 (cRT)	365	10,804	7,446	292	73	75.3%	Balanced val (seed=42)
GMDB	Qwen3.5-0.8B	508	849	1,724	406	102	51.6%	Original split
RareBench	Qwen3.5-0.8B (text)	508	117	78	—	—	42.3%	Random (seed=42)

For ImageNet-LT, we use the ERM ResNet-50 checkpoint from Kang et al. (2020) and split the 50,000-image ILSVRC 2012 validation set into 20,000 calibration and 30,000 test images (seed=42). For Places-LT, we extract frozen ResNet-152 features and retrain a logistic-regression classifier with class-balanced weighting (cRT); the Places365 balanced validation set (∼100 images per class, 18,250 total) is split 60/40 into calibration and test (stratified, seed=42). For RareBench, the GMDB-trained Qwen3.5-0.8B checkpoint is applied out-of-distribution to text-only clinical phenotype descriptions: HPO terms are formatted as a clinical prompt with a blank image placeholder, and the model outputs 508-dim logits over the GMDB disease label space. Of 1,122 RareBench cases, 195 map to 67 GMDB diseases; the remaining cases involve diseases outside the GMDB label set. The rare/frequent split for RareBench follows the GMDB training frequencies over the shared 508-class label space.

Figure 4 shows the training-set class frequency distribution for each benchmark, with the 80/20 rare/frequent boundary marked.

[1] https://github.com/facebookresearch/classifier-balancing
[2] https://db.gestaltmatcher.org
[3] https://huggingface.co/Qwen/Qwen3.5-0.8B
[4] https://github.com/chenxz1111/RareBench



(a) iNaturalist (b) ImageNet-LT (c) Places-LT

(d) GMDB (e) RareBench (f) Synthetic

Figure 4: Training-set class frequency distributions. Classes are sorted by frequency (descending); the vertical dashed line marks the 80/20 rare/frequent boundary.

F.2 Evaluation metrics

Hit@k. Hit@$k(\hat{y}) = P(\text{rank}_r(Y, X, S(X)) \leq k \mid C)$. Hit@1 is coverage-conditioned top-1 accuracy within the reranked shortlist. Rare Hit@1 and Frequent Hit@1 apply the same definition after restricting to the corresponding class subsets.

Mean Reciprocal Rank (MRR). $\text{MRR}(\hat{y}) = \mathbb{E}[\text{rank}_r(Y, X, S(X))^{-1} \mid C]$, where rank_r is the position of the true label in the reranked shortlist.

Recall@k. Recall@$k = P(Y \in S(X))$, the shortlist coverage. This is a property of the base model, not the reranker, and serves as the attainable ceiling for all reranking methods.

Recoverable gap closure (Eq. 9). $\rho_k = 1$ means the reranker closes the entire recoverable gap; $\rho_k = 0$ means no improvement.

Hardest-rival flip rate (Definition 6.1). Conditions on examples where the base model errs, so the metric is not inflated by examples already handled correctly.

G Additional Experiments

G.1 Synthetic experiment details

Data generation. We simulate $K = 100$ classes with a Zipf prior $\pi_y \propto 1/y$, generating $n_{\text{train}} = 5{,}000$ training and $n_{\text{test}} = 2{,}000$ test examples. Each class y has a fixed mean $\mu_y \sim \mathcal{N}(0, 0.5I)$ in $\mathbb{R}^{10}$. For each example, we sample $Y \sim \text{Categorical}(\pi)$ and $x \sim \mathcal{N}(\mu_Y, I)$. The base model score is $g_y(x) = \log p(y \mid x) + b_y + \epsilon$, where $b_y = -\tau \log \pi_y \cdot c$ is a class-level bias with corruption factor c, and $\epsilon \sim \mathcal{N}(0, 0.09)$. The shortlist is $S(x)$ = top-10 classes by $g(x)$.

Class-separable regime. We set $c = 0.1$, introducing a strong class-level bias without any context-dependent confusion. The residual $\beta^\star_y(x, S)$ is approximately constant per class up to a shortlist-shared constant.

Non-class-separable regime with contradictory pairs. We select 15 confuser pairs (u, v) where u is a tail class and v is a head class. For each example where both u and v appear in the shortlist, we perturb the base scores: $g_u(x) \mathrel{+}= s \cdot \delta$ and $g_v(x) \mathrel{-}= s \cdot \delta$, where $s \sim \text{Uniform}\{-1, +1\}$ is a per-example context sign and $\delta = 3.0$ is the confusion strength. This creates contradictory threshold requirements across contexts: $t(x; u, v)$ can be large positive in one context and large negative in another, which is exactly the condition of Theorem 4.3.

Fitting. The pairwise feature vector $\phi(x, y, j)$ consists of three features: the base score gap $g_y(x) - g_j(x)$, the rank gap, and the cosine similarity $\cos(\mu_y, \mu_j)$. We maximize the conditional log-likelihood (13) using L-BFGS with $\lambda_\theta = \lambda_a = 0.001$. All methods share the same shortlist; only covered examples ($y_i \in S_i$) are used for fitting. Results are averaged over 5 seeds.

G.2 Shrinkage diagnostic

For rare classes with only a handful of covered calibration examples, the unshrunk classwise estimate can have high variance. Figure 5 illustrates the intended behavior of the empirical-Bayes shrinkage in a controlled diagnostic: as calibration examples per class increase, the variance gap between the MLE offsets and the shrunk offsets narrows, while the tail-accuracy benefit of shrinkage is largest in the lowest-data regime.

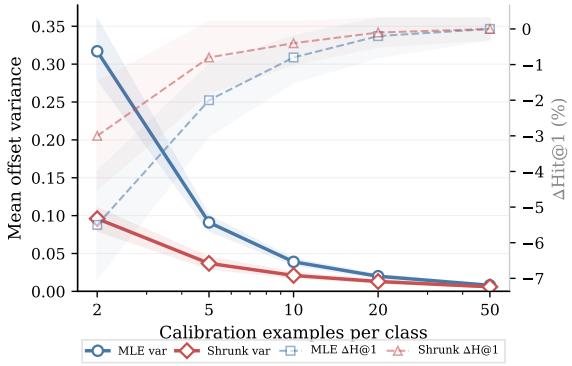


Figure 5: Shrinkage diagnostic for the classwise correction. Left axis: mean offset variance for the raw MLE and the shrunk estimate. Right axis: ΔHit@1. Shrinkage is most helpful when the number of covered calibration examples per class is very small.

G.3 Component ablation

Figure 6 shows the synthetic ablation in both regimes. Figure 7 extends the analysis to all five real datasets.

G.4 Real-data quintile analysis

Figure 8 shows the gain of REPAIR over Classwise stratified by mean per-class threshold dispersion D_y on the four real datasets.

RareBench is omitted because its small number of covered classes yields fewer than 15 classes per quintile, making the stratified analysis unreliable.

G.5 Sensitivity to shortlist size

Table 6 evaluates REPAIR across shortlist sizes $k \in \{5, 10, 20, 50\}$. REPAIR consistently outperforms all baselines at every k on every dataset. As k increases, coverage improves

Figure 6: Synthetic ablation (K=100, k=10). CW-only: θ=0 (classwise correction only). PW-only: a=0 (pairwise correction only). In the class-separable regime, CW-only matches REPAIR; in the non-class-separable regime, both components contribute.

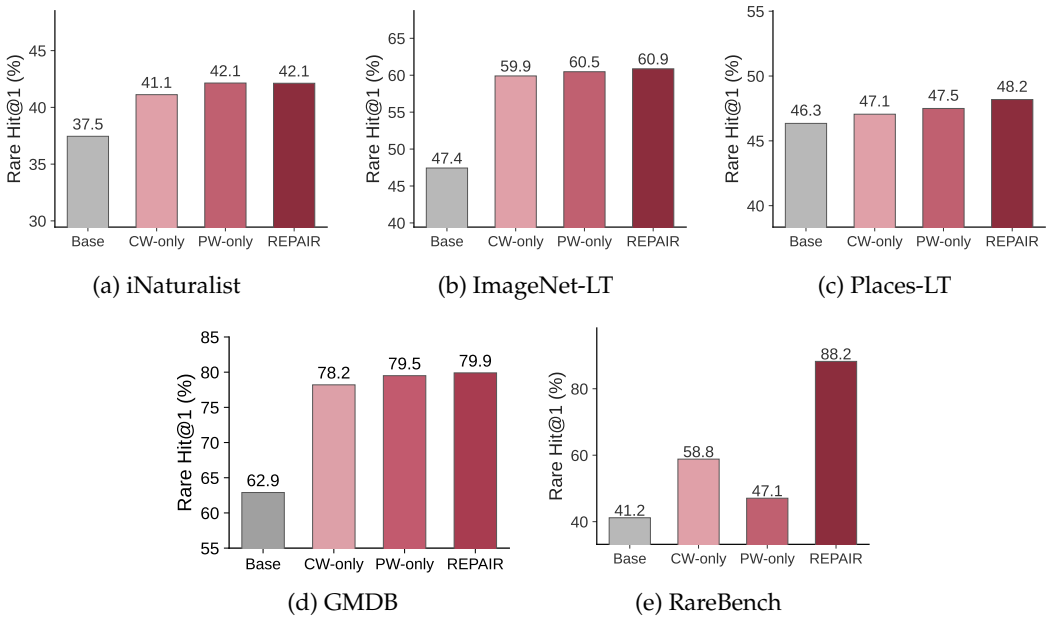


Figure 7: Component ablation on real datasets (Rare Hit@1). CW-only: $\theta = 0$. PW-only: $a = 0$. On iNaturalist, PW-only matches REPAIR (42.1); on GMDB, PW-only closely matches REPAIR (79.5 vs. 79.9). On ImageNet-LT and Places-LT, both components contribute incrementally. On RareBench, neither component alone approaches the full model (CW-only 58.8, PW-only 47.1, REPAIR 88.2).

but the reranking task becomes harder; the relative advantage of REPAIR over Classwise is stable or grows with k.

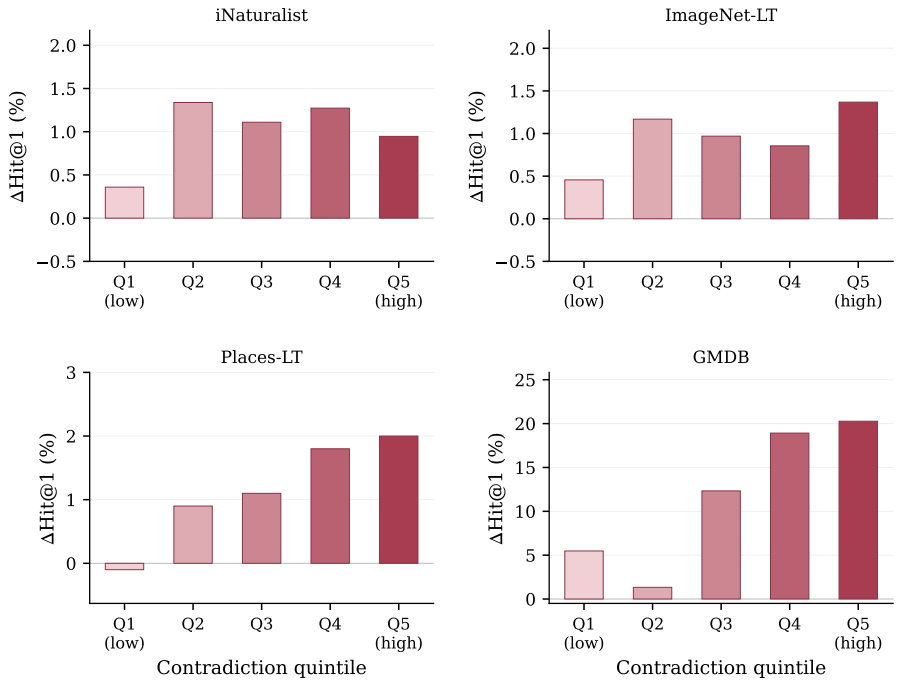


Figure 8: ΔHit@1 of REPAIR over Classwise by mean D_y quintile on real datasets. Classes are sorted by their mean threshold dispersion and divided into five equal-sized bins. Q5 (highest dispersion) shows the largest gain on ImageNet-LT, Places-LT, and GMDB. On iNaturalist, the overall gain is small ($<$ 1.5% per bin) and the quintile gradient is flat, consistent with a near-class-separable residual.

G.6 Unconditional results

Table 7 reports unconditional Hit@1, which factors in coverage: unconditional Hit@1 = conditional Hit@1 $\times$ Recall@k / 100. Since all methods share the same shortlist, unconditional differences mirror the conditional ones.

G.7 Sign test results

Table 8 reports the sign test results for REPAIR vs. Classwise across all dataset–metric pairs. Each cell shows the number of trials (out of 5) in which REPAIR strictly outperforms Classwise.

REPAIR significantly outperforms Classwise on Hit@1, Hit@3, MRR, and Rare Hit@1 across all five datasets (20/20 pass at $\geq$4/5). HFR passes on four of five datasets; RareBench (3/5) is limited by its small covered test set (32 examples). Frequent Hit@1 passes only on RareBench, where pairwise competition benefits both strata; on the remaining datasets the pairwise term introduces slight noise on frequent classes, as discussed in §6.2.

G.8 RareBench bootstrap confidence intervals

Because RareBench has only 32 covered test examples, we report bootstrap 95% confidence intervals (10,000 resamples) to assess the reliability of the observed gains.

The probability that REPAIR outperforms Classwise under resampling is $P = 0.999$. On rare classes specifically, the gap is +27.1% with 95% CI $[+11.8, +52.9]$ ($P = 0.998$). Despite the small sample size, the improvement is statistically robust.

Table 6: Sensitivity to shortlist size k. Hit@1 and Rare Hit@1 (%) conditioned on coverage. Best per dataset-k in **bold**.

		Hit@1 ↑				Rare Hit@1 ↑			
Dataset	Method	k=5	k=10	k=20	k=50	k=5	k=10	k=20	k=50
Synthetic (class-sep.)	Base	52.6	45.8	42.6	41.2	–	–	–	–
	Classwise	63.0	54.8	50.3	48.1	–	–	–	–
	REPAIR	**63.0**	**54.8**	**50.3**	**48.1**	–	–	–	–
Synthetic (non-class-sep.)	Base	59.0	52.6	49.6	48.4	–	–	–	–
	Classwise	68.1	61.1	57.2	55.2	–	–	–	–
	REPAIR	**79.5**	**72.5**	**68.3**	**66.1**	–	–	–	–
iNaturalist	Base	56.1	47.0	40.5	34.5	47.6	37.5	30.6	24.6
	LogitAdj	58.2	49.0	42.3	36.1	52.4	41.7	34.4	27.7
	Classwise	58.1	48.8	42.1	35.9	51.8	41.1	33.7	27.1
	REPAIR	**58.4**	**49.0**	**42.3**	**35.9**	**52.9**	**42.2**	**34.7**	**27.9**
ImageNet-LT	Base	61.6	54.4	49.5	45.7	55.2	47.4	42.4	38.5
	LogitAdj	69.2	61.5	56.0	51.7	68.4	59.5	53.2	48.4
	Classwise	69.5	61.8	56.3	51.9	68.9	59.9	53.6	48.7
	REPAIR	**69.7**	**62.0**	**56.4**	**52.0**	**69.7**	**60.8**	**54.4**	**49.3**
Places-LT	Base	52.4	44.7	39.5	35.8	54.0	46.3	41.0	37.2
	LogitAdj	52.1	44.5	39.3	35.6	55.8	47.9	42.3	38.5
	Classwise	52.9	45.1	39.9	36.0	54.8	46.9	41.6	37.6
	REPAIR	**53.4**	**45.5**	**39.9**	**36.2**	**56.2**	**48.2**	**42.4**	**38.5**
GMDB	Base	75.2	64.0	56.0	45.8	70.5	62.9	53.5	41.1
	LogitAdj	78.7	70.6	58.9	48.1	77.8	78.8	59.6	45.8
	Classwise	78.9	71.4	59.3	48.4	79.2	78.5	60.1	46.9
	REPAIR	**79.5**	**72.3**	**60.2**	**49.0**	**81.5**	**79.6**	**63.8**	**48.5**
RareBench	Base	62.1	59.4	50.0	37.5	46.7	41.2	38.9	25.9
	LogitAdj	65.5	62.5	52.8	39.6	53.3	52.9	44.4	29.6
	Classwise	72.4	66.2	72.2	56.2	60.0	54.1	77.8	55.6
	REPAIR	**86.2**	**85.6**	**88.9**	**72.9**	**80.0**	**81.2**	**88.9**	**70.4**

G.9 Pure-text NLP benchmarks

To test whether the residual decomposition holds beyond vision and clinical settings, we evaluate on four pure-text NLP benchmarks: GoEmotions (emotion classification, 28 classes), MASSIVE (intent classification, 60 classes), HuffPost-LT (news-topic classification, 42 classes), and Few-NERD (entity-type classification, 66 classes). All four use a Qwen3.5-2B base classifier fine-tuned with LoRA, and follow the same calibration/test protocol as the main experiments: reranker parameters are estimated on a held-out calibration split and evaluated on a disjoint test split, hyperparameters are selected on the calibration split by maximizing Hit@1 among covered examples, and all metrics are conditioned on coverage ($Y \in S$). We compare Base, LogitAdj, τ-norm, Classwise, and REPAIR under the identical protocol, with Classwise and REPAIR reporting mean$\pm$std over 5 calibration subsamples.

Table 10 reports the full comparison. The class-separable / non-class-separable distinction studied in the main text reproduces in pure-text LM classifiers. On GoEmotions, the residual is class-separable: Classwise already captures nearly all of the recoverable correction, and REPAIR adds only +0.05 Hit@1 and +0.51 Rare Hit@1 over it. On MASSIVE, HuffPost-LT, and Few-NERD, the residual is non-class-separable, and the pairwise term

Table 7: Unconditional results (k=10). Unconditional Hit@1 equals conditional Hit@1 scaled by Recall@10. Best per dataset in **bold**.

			Hit@1		Rare Hit@1	
Dataset	Method	Recall	Cond.	Uncond.	Cond.	Uncond.
iNaturalist	Base	37.8	47.0	17.8	37.5	14.2
	LogitAdj		49.0	18.5	41.7	15.8
	Classwise		48.8	18.5	41.1	15.6
	REPAIR		**49.0**	**18.5**	**42.2**	**15.9**
ImageNet-LT	Base	76.7	54.4	41.7	47.4	36.4
	LogitAdj		61.5	47.2	59.5	45.6
	Classwise		61.8	47.4	59.9	45.9
	REPAIR		**62.0**	**47.6**	**60.8**	**46.6**
Places-LT	Base	75.3	44.7	33.7	46.3	34.9
	LogitAdj		44.5	33.5	47.9	36.1
	Classwise		45.1	34.0	46.9	35.3
	REPAIR		**45.5**	**34.3**	**48.2**	**36.3**
GMDB	Base	51.6	64.0	33.0	62.9	32.5
	LogitAdj		70.6	36.4	78.8	40.7
	Classwise		71.4	36.8	78.5	40.5
	REPAIR		**72.3**	**37.3**	**79.6**	**41.1**
RareBench	Base	42.3	59.4	25.1	41.2	17.4
	LogitAdj		62.5	26.4	52.9	22.4
	Classwise		66.2	28.0	54.1	22.9
	REPAIR		**85.6**	**36.2**	**81.2**	**34.3**

Table 8: Sign test results (REPAIR vs. Classwise, 5 calibration subsamples). $\geq$4/5 wins = significant (✓).

Metric	iNat	ImageNet-LT	Places-LT	GMDB	RareBench
Hit@1	5/5 ✓	5/5 ✓	5/5 ✓	5/5 ✓	4/5 ✓
Hit@3	5/5 ✓	4/5 ✓	5/5 ✓	5/5 ✓	5/5 ✓
MRR	5/5 ✓	4/5 ✓	5/5 ✓	5/5 ✓	5/5 ✓
Rare Hit@1	5/5 ✓	5/5 ✓	5/5 ✓	5/5 ✓	4/5 ✓
Freq Hit@1	0/5	0/5	0/5	0/5	5/5 ✓
HFR	5/5 ✓	5/5 ✓	5/5 ✓	5/5 ✓	3/5

yields gains a fixed classwise offset cannot: +1.78, +0.84, and +1.05 Hit@1 over Classwise, with Rare Hit@1 gains of +5.53, +2.04, and +1.83, respectively. On FEW-NERD, LogitAdj improves Rare Hit@1 only modestly over Base (18.60 $\rightarrow$ 19.90), Classwise gives a larger improvement ($\rightarrow$ 29.75), and REPAIR improves it further ($\rightarrow$ 31.58), indicating that a learned classwise correction helps but the pairwise term still captures additional label-competition structure. This is consistent with the framework's prediction: pairwise correction helps precisely where the classwise component is insufficient.

G.10 HPO ablation on GMDB

A natural concern is whether REPAIR's gains in the rare-disease setting come from the HPO phenotype-similarity feature, an external source of medical knowledge, rather than from the pairwise correction itself. To test this, we remove the HPO Jaccard feature from $\phi(x, y, j)$ on GMDB while keeping the same calibration protocol and five calibration seeds. Table 11

Table 9: RareBench bootstrap 95% CI (n=32, 10K resamples). Even at the lower bound of the 95% CI, REPAIR outperforms Classwise by +6.2 %.

Method	Hit@1 (%)	95% CI
Base	59.4	[40.6, 75.0]
Classwise	66.2	[50.5, 81.8]
REPAIR	**85.6**	[73.1, 95.0]
REPAIR − Classwise	+19.4	[+6.2, +31.2]

Table 10: Reranking results on four pure-text NLP benchmarks (shortlist size 10), all using a Qwen3.5-2B base classifier fine-tuned with LoRA. All metrics are conditioned on the true label appearing in the shortlist; arrows indicate higher is better. Rare and Frequent denote the bottom 80% and top 20% of classes by training-set frequency. Both Classwise and REPAIR report mean±std over 5 calibration subsamples. Best per dataset in **bold**. Row shading: blue = class-separable, red = non-class-separable.

		Overall		*Stratified Hit@1*		
Dataset	Method	Hit@1↑	Hit@3↑	Rare↑	Freq.↑	HFR↑
GoEmotions (28 cls)	Base	58.10	81.23	26.90	74.97	–
	LogitAdj	58.48±.05	**81.58**±.16	29.97±.30	73.89±.12	.042
	τ-norm	58.27±.00	81.07±.00	27.21±.00	**75.08**±.00	.017
	Classwise	58.82±.07	81.54±.13	34.94±.81	71.93±.42	.093
	REPAIR (Ours)	**58.87**±.39	81.53±.07	**35.45**±2.95	71.92±.89	**.139**
MASSIVE (60 cls)	Base	55.26	77.65	32.06	68.55	–
	LogitAdj	56.39±.00	78.86±.00	40.07±.00	68.55±.00	.133
	τ-norm	55.26±.00	77.65±.00	32.06±.00	68.55±.00	.000
	Classwise	56.65±.16	78.71±.33	40.26±.55	67.70±.19	.122
	REPAIR (Ours)	**58.43**±.26	**80.79**±.20	**45.79**±.62	**71.30**±.61	**.217**
HuffPost-LT (42 cls)	Base	66.51	86.87	41.15	**81.90**	–
	LogitAdj	66.62±.04	87.19±.06	42.61±.11	81.20±.08	.026
	τ-norm	66.47±.08	86.82±.10	41.12±.05	81.85±.10	.002
	Classwise	66.14±.15	86.58±.14	43.39±1.69	80.60±.84	.060
	REPAIR (Ours)	**66.98**±.19	**87.53**±.18	**45.43**±2.33	80.07±1.14	**.062**
Few-NERD (66 cls)	Base	37.13	65.48	18.60	39.63	–
	LogitAdj	38.14±.00	67.34±.00	19.90±.00	40.43±.00	.132
	τ-norm	37.06±.03	65.30±.09	18.42±.09	39.57±.03	.001
	Classwise	38.64±.03	68.82±.14	29.75±.83	39.84±.09	.050
	REPAIR (Ours)	**39.69**±.03	**70.11**±.21	**31.58**±1.08	**40.78**±.13	**.137**

reports the result. Removing HPO changes every REPAIR metric by less than 0.1, well within seed standard deviation, so the aggregate gain is not driven by HPO specifically: the model-derived competition features (score gap, rank gap, log-probability ratio, log-frequency ratio) carry most of it. HPO remains useful qualitatively by shaping the composition of the top-k differential diagnosis list, as illustrated in Appendix G.11.

G.11 GMDB case study: the qualitative role of HPO

The HPO ablation in Appendix G.10 shows that HPO is not the main source of REPAIR's aggregate Hit@1 gain. Its contribution is instead qualitative: it shapes the composition of the top-k differential diagnosis list. We illustrate this with one held-out GMDB patient.

Table 11: GMDB ablation removing the HPO Jaccard feature (mean±std over 5 calibration subsamples). This ablation uses a separate calibration split, so absolute Hit@1 differs slightly from Table 1; the relevant comparison is REPAIR with HPO versus without HPO under the same split.

Method	Hit@1↑	Rare↑	Freq.↑	HFR↑
Classwise	$67.25_{\pm.68}$	$66.61_{\pm.74}$	$67.69_{\pm.92}$	.301
REPAIR with HPO	$70.55_{\pm.57}$	$76.12_{\pm.39}$	$66.78_{\pm1.11}$	.466
REPAIR without HPO	$70.50_{\pm.53}$	$76.06_{\pm.43}$	$66.75_{\pm1.07}$	.466

The patient presents with long palpebral fissures with eversion of the lateral lower eyelid, arched eyebrows, large protruding earlobes, mild intellectual disability, congenital heart defect, and short stature. The correct diagnosis is Kabuki syndrome. In the base Qwen3.5-0.8B shortlist, the clinically related CHARGE syndrome appears at rank 5 while the correct Kabuki syndrome appears at rank 8, so the case is recoverable but initially misranked. The two syndromes share overlapping facial, cardiac, hearing-related, and developmental phenotypes and are often part of the same clinical differential.

Table 12 shows the selected ranks after reranking. A fixed frequency-based offset (LogitAdj) can lift Kabuki syndrome to rank 1, but it pushes CHARGE syndrome down to rank 7: it corrects the top-1 label while disrupting a clinically meaningful differential. REPAIR with HPO similarity also promotes Kabuki syndrome to rank 1, but keeps CHARGE syndrome at rank 2. Because Kabuki and CHARGE share relevant phenotype terms, the HPO-aware pairwise term gives their pair a stronger positive correction than unrelated rare diseases receive. Thus the pairwise HPO term does not merely improve top-1 accuracy; it changes the structure of the top-k list, lifting the correct diagnosis while preserving a phenotypically related alternative near the top.

Table 12: Selected ranks for one held-out GMDB patient (correct diagnosis: Kabuki syndrome). CHARGE syndrome is a clinically related differential. LogitAdj corrects the top-1 prediction but pushes the related diagnosis down the list, whereas REPAIR with HPO similarity corrects the top-1 prediction while keeping the related diagnosis near the top.

Method	Kabuki (correct)	CHARGE (related)
Base Qwen3.5-0.8B	Rank 8	Rank 5
LogitAdj	Rank 1	Rank 7
REPAIR with HPO similarity	Rank 1	Rank 2

H Limitations

Domain-specific features. The pairwise feature vector includes a domain-specific similarity component (taxonomic distance, WordNet similarity, or HPO Jaccard) on four of five datasets. When such structured knowledge is unavailable, as on Places-LT, we fall back to a four-dimensional feature vector without domain similarity and still observe gains. A natural alternative in the absence of external knowledge is cosine similarity between the base model's class embeddings, which we leave for future work.

Shortlist coverage ceiling. REPAIR operates on a fixed top-k shortlist produced by the base model and cannot recover examples where the true label falls outside the shortlist. This is a fundamental constraint shared by all post-hoc rerankers, not specific to our method. Our framework explicitly conditions on coverage ($Y \in S$) and focuses on maximizing ranking quality within this recoverable set. Improving coverage itself requires modifying the base model, which is orthogonal to the reranking problem we study.