EVALUATION COLLAPSE: WHEN METRICS CANNOT DISTINGUISH MODEL BEHAVIOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Scalar evaluation metrics (accuracy, AUROC, F1, perplexity) rest on three structural restrictions: pointwise scoring, permutation invariance, scalar aggregation. We prove this combination cannot represent behavioral properties—detection delay, subgroup performance, confidence trajectories—that depend on discarded information. For any such property, $\Theta(n)$ behaviorally distinct models receive identical metric scores, even at the maximum score where practitioners are most confident. No metric can recover what its representation discards. We prove: (1) collapse occurs whenever behavioral information is absent; (2) collapse families persist at optimal scores, making metric tuning futile; (3) for ordered sequences, collapse size follows an exact combinatorial law (Theorem 3). We instantiate this in three settings: anomaly detectors at AUROC=1.000 differing by 50 steps in latency, classifiers with identical accuracy differing by 20 points in subgroup error rates, and language models with identical perplexity spanning full confidence-trajectory ranges. The phenomenon appears in real systems: COMPAS, a deployed recidivism-risk tool, exhibits this signature—near-identical accuracy across race concealing 20-point divergence in false-positive/negative rates—independently reproduced from public data. We operationalize the result as a lightweight audit on already-computed scores (no retraining), with characterized failure modes: it reliably detects collapse but can underestimate by 50× if perturbation distribution is too narrow. The core question: before certifying behavior from a metric, does the evaluation retain enough information to see what you're certifying?

1 INTRODUCTION

Machine learning models are often evaluated with a single number. Accuracy, AUROC, F1, and perplexity are widely used to compare models, select checkpoints, and guide deployment decisions. A scalar score is useful because it makes comparison simple. But reducing model behavior to one number also removes information, and this is not a hypothetical concern.

In 2016, ProPublica's investigation of the COMPAS recidivism-risk tool, used in pretrial and sentencing decisions across the United States, found that the algorithm's overall accuracy was nearly identical across race (Angwin et al., 2016). The vendor defended the tool as fair on the strength of this aggregate parity. But the near-identical accuracy concealed a stark divergence in *how* the two groups were misclassified: one group was substantially more likely to be incorrectly flagged as high-risk, the other substantially more likely to be incorrectly cleared as low-risk. The subsequent public dispute was in large part a disagreement about which projection of a single collapsed statistic to trust: each side's claim was consistent with the reported score but invisible to it individually (Section 3.7.3).

This is consistent with the phenomenon our framework formalizes: accuracy, computed pointwise and aggregated without regard to subgroup membership, is structurally incapable of representing the subgroup error asymmetry that made COMPAS controversial, because the aggregation discards the information needed to see it.

The same structural blindness recurs elsewhere. Consider two clinical anomaly detectors monitoring patient physiological data for signs of sepsis, both achieving AUROC $= 1.000$, so every anomalous observation receives a higher score than every normal observation. The models can nevertheless behave very differently over time: one may raise an alert at the first anomalous step, while the other may require 50 additional observations. This difference can determine whether an intervention occurs

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.

in time, yet AUROC cannot distinguish the models, since it depends only on the relative ordering of scores, not on when alerts occur. A gap of exactly this kind is implicated in the Epic Sepsis Model, a widely deployed clinical prediction tool: independent external validation found an AUC of 0.63, substantially below the 0.76–0.83 range reported internally by the vendor (Wong et al., 2021), consistent with Theorem 1's prediction that detection latency can diverge sharply while an aggregate discrimination score remains high in development but degrades in deployment.

The same issue appears in language modeling. Two language models can have identical or near-identical perplexity while diverging sharply in their confidence trajectories, that is, how per-token confidence evolves over a sequence, a disconnect independently documented across recent studies relating perplexity to downstream task performance (Chen et al., 2024). In each case, the reported score can be correct while failing to reveal a behavior that matters for use.

These failures are usually handled separately. Fairness evaluations report subgroup performance, anomaly detection uses delay-sensitive measures, and language-model evaluation uses measures of calibration or downstream performance. Such methods are useful, but they leave a more basic question open: when is a behavioral property impossible to recover from a given evaluation score in the first place?

We study this question as a problem of information loss. We consider evaluation methods that compute a score for each example, ignore the order of those scores, and reduce them to one number. We show that when a behavioral property is not recorded by this process, no choice of the final scoring rule can recover it. We construct families of $\Theta(n)$ models with different behavior that receive exactly the same evaluation score, a phenomenon that persists at the maximum attainable score, and for the ordered score sequence itself is not a bound but an exact orbit count.

The result is therefore about the structure of the evaluation, not a weakness of any particular metric. AUROC, accuracy, F1, and perplexity are examples of this broader class; the same argument applies to any scalar evaluation method that discards the information needed to distinguish the behavior of interest.

Our theory has four parts. First, missing behavioral information produces exact score collapse. Second, the number of behaviorally different models sharing a score can grow linearly, or exponentially in the language-modeling case, with the number of evaluation examples, and can persist at the maximum attainable score. Third, for the ordered score sequence itself, we give an exact combinatorial count of the collapse family via tie-orbit counting, rather than an asymptotic bound. Fourth, each of the three structural conditions can be individually relaxed to recover the corresponding behavioral information, which we operationalize as a runnable diagnostic (Section 4).

We demonstrate the result in three settings, including one real, deployed, publicly scrutinized system. In clinical anomaly detection, two models achieve AUROC $= 1.000$ while differing by 50 steps in detection delay. In classification, models with identical accuracy and AUROC exhibit subgroup performance gaps of up to 20 points, and we show this same signature, near-matched accuracy alongside large subgroup error divergence, in the COMPAS dataset itself. In language modeling, models with identical perplexity span the full range of possible confidence trajectories, an exact instance of our tie-orbit result and a pattern we also confirm on real GPT-2 log-probabilities.

Our contributions are:

- **Structural Impossibility Theorem.** An evaluation method cannot distinguish behavioral properties absent from the information it uses.
- **Quantitative Collapse Bounds.** Collapse families of size $\Theta(n)$ or $\Theta(2^T/\sqrt{T})$, including at the maximum attainable score.
- **An Exact Collapse Law for Tied Scores.** For the ordered score sequence itself, the collapse-family size is exactly $n!/\prod_j n_j!$ by orbit-stabilizer counting (Theorem 3), of which the perplexity construction is a direct instance.
- **Sufficiency of the Conditions.** Retaining order, group information, or the full output value is individually sufficient to recover the corresponding behavior.
- **A Runnable Diagnostic.** A lightweight audit, applicable to already-computed scores, that estimates whether a reported score is evidence about a given behavioral property, with characterized false-negative failure modes.

- **Real-World Grounding.** The phenomenon is independently reproducible in a real deployed system (COMPAS), not merely constructible.

2 Related Work

Impossibility results in evaluation. Our theorem sits in a lineage of formal impossibility results showing that desiderata for a summary statistic cannot simultaneously hold. Kleinberg et al. (2016) and Chouldechova (2017) show that calibration and error-rate balance cannot generally be satisfied simultaneously. These concern tension *between* multiple target properties of a fixed statistic; ours is different in kind: a *single* scalar statistic is structurally incapable of representing any property outside a characterizable information class, regardless of how it is chosen or tuned. This maps onto classical *sufficiency*: a pointwise, permutation-invariant, scalar reduction of $(s_1, \ldots, s_n)$ is a lossy statistic relative to the full order statistic, and by a Fisher–Neyman-style argument, any property not measurable from it is provably unrecoverable. We quantify this loss ($\Theta(n)$, $\Theta(2^T/\sqrt{T})$, or exact via Theorem 3) and identify which relaxation restores which property.

Domain-specific remedies as instances of our propositions. The Numenta Anomaly Benchmark (Lavin & Ahmad, 2015) penalizes late detections directly, and sequential/online protocols track cumulative regret rather than a terminal score, an instance of Proposition 1. Subgroup accuracy reporting (Buolamwini & Gebru, 2018) is the applied analogue of Proposition 2. Both are effective, domain-specific responses to the same structural cause; we reframe them as instances of a single general theorem, and give a closed-form bound (subgroup case) on how much disparity a fixed accuracy score can hide.

Calibration. Expected Calibration Error and reliability diagrams (Guo et al., 2017) retain a structured, non-scalar representation of confidence, an instance of Proposition 3. Our perplexity/trajectory construction shows precisely what a scalar log-likelihood score discards that ECE-style statistics recover.

What is new here. Prior work treats delay-sensitivity, fairness disaggregation, and calibration as three separate remedies to three separate problems. We show they are the same remedy applied along three independent, formally distinguished axes of one structural cause. Unlike adversarial metric gaming, an instance of Goodhart's Law where perverse incentives degrade unmeasured behavior, collapse occurs even under good-faith evaluation, and the Collapse Audit (Section 4) defends against both. We give quantitative (linear/exponential) collapse bounds and a maximum-score result beyond simple non-identifiability; show that for tied scores the collapse-family size is not merely bounded but an exact combinatorial law (Theorem 3); and turn the theorem into a runnable audit and a confirmed real-world instance (Section 3.7.3) rather than leaving the loss abstract.

3 Framework and Main Theorem

3.1 Setup and Behavioral Identifiability

Let a model f produce outputs on an evaluation dataset $D = \{(x_i, y_i)\}_{i=1}^n$. A *behavioral dimension* $\phi(f, D)$ is any property of the model and dataset, for example the time of first alert, the vector of per-group accuracies, or the sequence of token-level confidence scores. An *evaluation representation* $\mathcal{E}(f, D)$ is the observable summary an evaluation computes. A *metric* $\mathcal{M}(f, D) = G(\mathcal{E}(f, D))$ applies a scoring rule G to that representation, so $\mathcal{M}$ depends on (f, D) only through $\mathcal{E}$. A behavioral dimension ϕ is *determined by* $\mathcal{E}$ if $\phi(f, D) = h(\mathcal{E}(f, D))$ for some function h and all (f, D); otherwise there exist configurations $(f, D), (f', D')$ with $\mathcal{E}(f, D) = \mathcal{E}(f', D')$ but $\phi(f, D) \neq \phi(f', D')$. A *collapse family* $\mathcal{C}_r$ at metric value r is the set of configurations (f, D) with $\mathcal{M}(f, D) = r$ whose ϕ-values are not all identical.

3.2 Foundational Lemma

Lemma 1 (Information-Theoretic Impossibility)**.** *If a behavioral dimension ϕ is not determined by an evaluation representation $\mathcal{E}$, then for any metric $\mathcal{M} = G \circ \mathcal{E}$, there exist (f, D), (f', D') with $\mathcal{E}(f, D) = \mathcal{E}(f', D')$ but $\phi(f, D) \neq \phi(f', D')$, and hence $\mathcal{M}(f, D) = \mathcal{M}(f', D')$.*

Proof. Non-determinism of ϕ gives such a pair by definition. Since G is a function, applying it to both sides preserves the equality. ✓

This holds regardless of whether $\mathcal{E}$ is scalar, pointwise, or ordered. It is the seed from which Theorems 1 through 3 derive quantitative collapse. The three properties below explain why behavioral information becomes absent in standard metrics.

3.3 Three Structural Properties

Definition 1 (Pointwise Evaluation)**.** *$\mathcal{E}(f, D) = H(s_1, \ldots, s_n)$, where $s_i = s(f(x_i), y_i)$ depends only on the model output and label for instance i, with no access to group membership or other instances.*

Definition 2 (Permutation Invariance)**.** *H is symmetric under reordering: $H(s_1, \ldots, s_n) = H(s_{\sigma(1)}, \ldots, s_{\sigma(n)})$ for any permutation σ of $[n]$.*

Definition 3 (Scalar Aggregation)**.** *$\mathcal{E}(f, D) \in \mathbb{R}$: the representation retains no vectors, sequences, or curves.*

Standard metrics, including accuracy, AUROC, F1, and perplexity, satisfy all three. The properties are formally independent: pointwise evaluation restricts how each score is computed, permutation invariance restricts how scores are aggregated, and scalar aggregation restricts the output's form. Relaxing any one recovers different behavioral information: order recovers temporal behavior, instance metadata recovers group behavior, and the full sequence recovers trajectory behavior (Propositions 1 through 3).

3.4 Theorem 1: Linear Collapse from Structural Properties

Theorem 1 (Linear Collapse under Permutation Invariance)**.** *Let $\mathcal{E}$ satisfy pointwise evaluation and permutation invariance. Fix f, D. Suppose there exist $k_n = \Theta(n)$ permutations $\sigma_1, \ldots, \sigma_{k_n}$ of $[n]$ such that $\phi(f, \sigma_i D) \neq \phi(f, \sigma_j D)$ for all $i \neq j$. Then for any metric $\mathcal{M} = G \circ \mathcal{E}$, the collapse family at $r = \mathcal{M}(f, D)$ contains $\Theta(n)$ behaviorally distinct configurations.*

Proof sketch. Permutation invariance gives $\mathcal{E}(f, \sigma_j D) = \mathcal{E}(f, D)$ for every j, since each σ_j only reorders the score multiset. Applying G then gives $\mathcal{M}(f, \sigma_j D) = r$ for all j. Since ϕ is pairwise distinct across the σ_j by hypothesis, all $k_n = \Theta(n)$ configurations are behaviorally distinct. Full proof in Appendix A. ✓

Remark 1 (Tightness)**.** *The $\Theta(n)$ rate is tight given only these two properties: a single collision, guaranteed by Lemma 1, is forced by permutation invariance to propagate to an entire linear-size orbit, but not further without additional structure. Reaching a faster, super-linear rate, as Theorem 3 does below, requires more than permutation invariance alone.*

3.5 Theorem 2: Collapse at Optimal Scores

Definition 4 (Optimal-Score Family)**.** *A collapse family from Theorem 1 is* optimal *if some configuration attains $\max_{f,D} \mathcal{M}(f, D)$.*

Theorem 2 (Collapse at Maximum Score)**.** *If a linear collapse family is optimal-score, the collapse family at $r^* = \max_{f,D} \mathcal{M}(f, D)$ contains $\Theta(n)$ behaviorally distinct configurations.*

Proof sketch. If one configuration attains r^*, permutation invariance forces every configuration in its orbit to share the same $\mathcal{E}$-value, and hence the same $\mathcal{M}$-value r^*. Since ϕ remains pairwise distinct across the orbit, all $\Theta(n)$ configurations collapse at the maximum. Full proof in Appendix A. ✓

Corollary 1 (Optimality is Not Self-Correcting)**.** *No choice of scoring rule G can shrink an optimal-score collapse family: every configuration in the orbit shares one representation value, so any G maps the whole orbit to a single score. Only relaxing $\mathcal{E}$ itself, via Propositions 1 through 3, restores identifiability at the optimum. Full proof in Appendix A.*

Theorem 2 and Corollary 1 together show that metric blindness can persist exactly where it is least expected: at a perfect or near-perfect score, where a practitioner has the strongest incentive to treat the result as settled, no adjustment to the scoring function, however sophisticated, can recover the

missing behavioral distinction. The fix must come from the evaluation representation, not the metric built on top of it.

3.6 Theorem 3: Exact Collapse via Tie-Orbit Counting

Theorems 1 and 2 bound collapse-family size from below for an arbitrary behavioral dimension ϕ, via a constructed subfamily of permutations. We now show that for the ordered score sequence itself, no construction is needed: the collapse-family size is not a bound at all, but an exact group-theoretic invariant of the score sequence, namely the size of its orbit under the full symmetric group.

Theorem 3 (Exact Orbit Collapse)**.** *Let $\mathcal{E}$ satisfy pointwise evaluation and permutation invariance, and let s be the score sequence of (f, D) with tie multiplicities $n_1, \ldots, n_k$. Then, for the ordered score sequence $\phi_{\mathrm{ord}}(f, D) \coloneqq (s_1, \ldots, s_n)$, the collapse family at $r = \mathcal{M}(f, D)$ contains* exactly

$$\frac{n!}{n_1!\, n_2! \cdots n_k!}$$

behaviorally distinct configurations under ϕ_{ord}: the exact orbit-stabilizer count of S_n acting on s, not merely a lower bound.

Proof sketch. The stabilizer of s under coordinate permutation is $S_{n_1} \times \cdots \times S_{n_k}$, since permutations moving indices only within a tie class leave s unchanged, while any permutation crossing tie classes does not. This stabilizer has order $\prod_j n_j!$, giving orbit size $n!/\prod_j n_j!$ by the orbit-stabilizer theorem. Every $\sigma \in S_n$ leaves $\mathcal{M}$ unchanged, by Theorem 1's argument applied to the full group rather than a constructed subfamily, and two permutations yield the same ϕ_{ord}-value if and only if they lie in the same stabilizer coset. So the count of behaviorally distinct configurations under ϕ_{ord} equals the orbit size exactly, neither more nor fewer. Full proof and the score-sequence orbit definition in Appendix A.5. ✓

Remark 2 (Scope and Instance)**.** *Exactness holds for ϕ_{ord} and any injective function of it, including the perplexity and trajectory construction (Section 3.7.4), where $\phi_{\mathrm{traj}} = \phi_{\mathrm{ord}}$ directly, giving collapse size exactly $\binom{T}{m} = \Theta(2^T/\sqrt{T})$ at $m = T/2$ (Appendix A.7). It does* not *extend to many-to-one statistics such as detection latency or subgroup disparity, which remain covered only by the asymptotic bounds of Theorems 1 and 2 (Appendix A.6). Nor is large-orbit collapse restricted to constructed datasets: any evaluation with a fixed, finite number of representable score levels generically produces exponentially large tie orbits as n grows (Proposition 4, Appendix A.8).*

3.7 Explicit Constructions

Table 1 summarizes the three instantiations, including their real-world counterparts. Full derivations for each are given in Appendix B.

Table 1: Explicit constructions and their properties.

Setting	Metric	Collapse Size	Optimal?	Real-world instance
Anomaly detection	AUROC	$\Theta(T)$	Yes	Epic Sepsis Model (Wong et al., 2021)
Subgroup evaluation	Accuracy	$\Theta(n)$	No	COMPAS (Angwin et al., 2016)
Language modeling	Perplexity	$\Theta(2^T/\sqrt{T})$	No	Pretraining and evaluation studies (Chen et al., 2024)

3.7.1 AUROC and Detection Latency (Optimal-Score Collapse)

For T time steps with anomaly onset t_0, we construct $\Theta(T)$ detectors, one per candidate alert position $t \in \{t_0, \ldots, t_0 + (T - t_0)/2\}$, each achieving AUROC $= 1.000$ while detection latency ranges from 0 to $(T - t_0)/2$ steps (Appendix B): an optimal-score instance of Theorem 2. On real NAB machine-temperature data, a trained detector achieves AUROC $= 0.830$ at zero simulated latency, degrading smoothly as latency increases (Lavin & Ahmad, 2015) (full curve in our released notebook).

3.7.2 Overall Accuracy and Subgroup Disparity (Fixed-Score Collapse)

For two groups of size m with a fixed total of $k = \Theta(m)$ errors, all $\Theta(m)$ error allocations between groups share the same overall accuracy $\mathrm{Acc} = 1 - k/(2m)$ while subgroup disparity ranges from 0 to k/m (Appendix B): a non-optimal instance of Theorem 1.

3.7.3 Real-World Instance: COMPAS and the ProPublica-Northpointe Dispute

The construction above is not merely illustrative: a close real-world analogue is documented in the COMPAS recidivism-risk controversy (Angwin et al., 2016). We independently reproduce the relevant statistics from ProPublica's public dataset using their published filtering criteria (Table 2).

Table 2: COMPAS accuracy and error rates by race, independently reproduced from ProPublica's public dataset (Angwin et al., 2016). Values are consistent with, though not identical to, the originally published figures due to minor differences in dataset snapshot and edge-case filtering.

Group	Overall Accuracy	FPR	FNR
White defendants	67.2%	22.0%	49.6%
Black defendants	64.9%	42.3%	28.5%
Gap	2.3 pts	20.3 pts	21.2 pts

Two subpopulations, one classifier: near-identical overall accuracy, a 2.3-point gap well within finite-sample noise, alongside large, oppositely-signed subgroup error disparities, a 20.3-point FPR gap and a 21.2-point FNR gap. This is consistent with, rather than a direct instance of, the collapse mechanism of Section 3: accuracy is a pointwise, scalar statistic that does not retain group membership, so by Lemma 1 it cannot determine subgroup accuracy for any population where the two are not perfectly correlated, independent of whether that population was constructed to exhibit the full $\Theta(n)$ collapse family of Theorem 1. COMPAS is real-world evidence that this structural blindness is not merely a worst-case possibility but occurs in practice, not a claim that the COMPAS population instantiates our synthetic construction. The tool's vendor's defense, predictive parity, and ProPublica's critique, error-rate imbalance, are both correct descriptions of the same score: they read off different projections of a representation, $\mathcal{E} =$ accuracy, that individually determines neither. Proposition 2 gives the resolution: retaining group membership alongside per-instance scores recovers exactly the information a single accuracy number discards.

We emphasize what this case study does *not* show. It does not establish bad faith, nor adjudicate which fairness criterion, error-rate balance or predictive parity, is normatively correct; that tension is a separate, well-studied impossibility result (Kleinberg et al., 2016; Chouldechova, 2017), orthogonal to ours. It shows that the descriptive question, does the accuracy number tell us what we need to know, was answered incorrectly by relying on the scalar figure alone, exactly as Theorem 1 predicts.

3.7.4 Perplexity and Confidence Trajectory (Fixed-Score Collapse)

For T tokens with m log-probabilities ℓ_1 and $T - m$ log-probabilities $\ell_2 > \ell_1$, every permutation of positions yields identical perplexity while defining $\binom{T}{m} = \Theta(2^T/\sqrt{T})$, at $m = T/2$, behaviorally distinct confidence trajectories $\phi_{\text{traj}} = (\ell_{\sigma(1)}, \ldots, \ell_{\sigma(T)})$. Since $\phi_{\text{traj}} = \phi_{\text{ord}}$ by definition, this is not merely an asymptotic instance of Theorem 1 but an *exact* instance of Theorem 3 (Remark 2; derivation in Appendix A.7), confirmed on real GPT-2 log-probabilities in our released notebook.

3.8 Sufficiency: Recovering Behavior by Retaining Information

Each structural restriction can be individually relaxed to recover the corresponding behavioral dimension.

Proposition 1 (Retaining Temporal Order)**.** *If $\mathcal{E}'(f, D) = (s(f(x_1), y_1), \ldots, s(f(x_n), y_n))$ retains order, detection latency $\phi_{\text{latency}} = \min\{t : s(f(x_t), y_t) > \tau\}$ is determined by $\mathcal{E}'$.*

Proposition 2 (Retaining Group Membership)**.** *If $\mathcal{E}'(f, D) = H((s_1, g_1), \ldots, (s_n, g_n))$ retains group labels, subgroup accuracies $\phi_{\text{sub}} = (\mathrm{Acc}(f|_{G_1}), \ldots, \mathrm{Acc}(f|_{G_k}))$ are determined by $\mathcal{E}'$.*

Proposition 3 (Retaining the Full Score Sequence)**.** *If $\mathcal{E}'(f, D) = (s_1, \ldots, s_n)$ retains the full sequence, the trajectory $\phi_{traj} = (\ell_1, \ldots, \ell_n)$ is determined by $\mathcal{E}'$.*

Each proposition shows that relaxing one restriction is sufficient to recover the corresponding behavior: retaining order recovers temporal behavior, group labels recover subgroup behavior, and the full sequence recovers trajectory behavior. We do not claim these relaxations are the only way to recover each behavior; we leave characterizing full necessity, whether some other single relaxation could substitute, to future work.

4 The Collapse Audit: A Practical Diagnostic

Theorems 1 and 2 are existential: they guarantee a collapse family exists whenever the structural conditions hold. In practice, a practitioner does not need to enumerate that family; they need to know, for their metric and their specific model and dataset, whether the reported score is doing the work they think it is. We give a lightweight audit procedure that estimates this directly from the per-instance scores already computed during evaluation, without retraining any model.

Algorithm 1 Collapse Audit: Detect Metric Blindness from Scores

Require: per-instance scores $s_1, \ldots, s_n$, labels $y_1, \ldots, y_n$, metric $\mathcal{M} = G \circ \mathcal{E}$, candidate behavioral dimension ϕ, resampling budget B

$r \leftarrow \mathcal{M}(s_1, \ldots, s_n)$ ▷ observed score
$\Phi \leftarrow \emptyset$
for $b = 1$ to B **do**
    Draw a structure-preserving perturbation σ_b (a permutation for order-sensitivity, a group-reassignment for subgroup-sensitivity, or a score-preserving relabeling for trajectory-sensitivity)
    if $\mathcal{E}(\sigma_b \cdot s) \neq \mathcal{E}(s)$ **then continue**
    end if
    $\Phi \leftarrow \Phi \cup \{\phi(\sigma_b \cdot s)\}$
end for
return spread(Φ), $|\Phi_{\text{distinct}}|$, and convergence status

Interpretation and cost. A large spread in ϕ across metric-invisible perturbations means the reported score r is *not* evidence about ϕ: all values in Φ are equally consistent with r. A small or zero spread means either the metric is not blind to ϕ for this dataset, or the audit missed collapse due to limited budget or a narrow perturbation distribution (Section 4.2). Each round is $O(n)$ and requires no retraining, so it is cheap to run alongside any leaderboard or model card. Use the widest perturbation distribution computationally feasible; start with $B = 100$ and confirm convergence by rerunning at increasing B until spread stabilizes.

Empirical validation. Table 3 applies Algorithm 1 to the three constructions of Section 3 with $B = 3000$. Empirical spread matches the closed-form bound almost exactly for AUROC/latency and accuracy/disparity; for the trajectory construction it reaches a substantial but not maximal fraction of the theoretical range, since a single random shuffle rarely lands on the fully sorted extreme.

Table 3: Collapse Audit results on the three constructions of Section 3 ($B = 3000$).

Setting	Observed score	Metric-invisible	Empirical spread	Theoretical spread
Anomaly detection (AUROC)	1.000000	3000/3000	50.0 steps	50 steps
Subgroup evaluation (Accuracy)	0.900000	3000/3000	0.2000 disparity	0.2000 disparity
Language modeling (Perplexity)	3.490343	3000/3000	1.4807 (ρ)	2.0 (max)

4.1 Robustness of the Constructions

The constructions in Section 3 use hand-placed score bands for clarity; we confirm the collapse signature is unchanged under continuous, randomly-drawn scores (Beta-distributed within each band) and across independent random seeds (20 for accuracy/disparity, 10 for perplexity/trajectory), with

disparity and perplexity curves overlapping to within floating-point error in every case. This confirms collapse is a structural property of the metric, not an artifact of a particular construction or draw; full sweeps and plots are in Appendix C and the released notebook.

4.2 Failure Modes and False Negatives

The Collapse Audit's guarantees run in only one direction: a large empirical spread is reliable positive evidence of collapse. A small or zero spread is *not* reliable negative evidence, since it can arise from an inadequate audit rather than an absent collapse family. Two distinct failure modes occur.

Budget too low. With too few resampling rounds B, the audit can miss rare metric-invisible perturbations even when a large collapse family exists. On the subgroup-disparity construction (true spread 0.20), empirical spread reaches only 50% of the true value at $B = 3$, 98% by $B = 25$, and 100% by $B = 50$. This failure mode is benign: it resolves with more compute and is detectable by checking convergence across increasing B.

Perturbation distribution too narrow. This failure mode is serious and does not resolve with a larger budget. If σ_b explores only a local neighborhood of the observed configuration, rather than the full space of metric-invisible alternatives, the audit systematically underestimates true collapse-family size regardless of B. Restricting the subgroup-disparity audit to perturbations within radius r of the observed error allocation, rather than the full range $[0, k]$, produces empirical spread of only 2% of the true value at $r = 1$ and 50% at $r = 25$, even at $B = 3000$; the same pattern holds on the AUROC/latency construction, with spreads of 4%, 8%, 20%, and 40% at $r \in \{1, 2, 5, 10\}$, reaching 100% only once r spans the full latency range. A practitioner who audits with a "natural" local perturbation, such as swapping only adjacent instances or reassigning labels only to nearby groups, can confidently observe a low spread and wrongly conclude the metric is not blind, when the true collapse family is large but lies outside the region the audit explored.

Practical implication. A zero-spread or small-spread result should be read as "no collapse was found by this audit, with this perturbation distribution and budget," not "no collapse exists." Defending against the second failure mode requires the widest feasible perturbation distribution and confirming that spread has converged across increasing B, not a single audit at fixed settings. Full budget- and radius-sweep ablations are in Appendix F.

5 Discussion

When does collapse actually matter? Not every use of a scalar metric is at risk. Collapse is a problem precisely when a decision downstream of the score depends on a behavioral dimension the score cannot see: selecting a sepsis detector based on AUROC alone, certifying a hiring or lending classifier as fair based on accuracy alone, or ranking language models by perplexity when deployment is sensitive to confidence calibration rather than average log-likelihood. The failure mode is always the same: a decision-maker treats $\mathcal{M}(f, D) = \mathcal{M}(f', D')$ as evidence that f and f' are behaviorally equivalent, when Theorems 1–2 guarantee this is unsound whenever ϕ is not determined by $\mathcal{E}$. Conversely, when downstream decisions do not depend on order, subgroup identity, or trajectory shape, a scalar summary is exactly the right tool. Collapse is a property of the mismatch between metric structure and decision requirements, not an indictment of scalars in general. Applying this diagnosis requires no new metrics, benchmarks, or retraining, only a check of what the pipeline already discards, and is cheapest exactly where it matters most: at the tied or near-tied scores where a leaderboard or model selection would otherwise proceed on informal grounds.

Practical protocol. Before certifying a behavioral claim from a metric: (1) *identify* the behavioral dimension ϕ the decision depends on, e.g. detection latency, subgroup accuracy, or confidence trajectory; (2) *check* whether $\mathcal{E}$ retains the relevant information, using Propositions 1–3 as a guide (order for timing, group labels for cohort effects, the full sequence for trajectory shape); and (3) *audit* when in doubt, running Algorithm 1 directly on stored scores, reporting a null result as "no collapse found by this audit" rather than "no collapse exists."

Table 5 (Appendix G) summarizes which structural relaxation addresses each behavioral concern; the same prescriptions apply to F1, top-k accuracy, BLEU, or any pointwise, order-blind, scalar

metric. Where ϕ is the ordered score sequence itself, Theorem 3 sharpens the $\Theta(n)$ bound to an exact orbit-stabilizer count.

6 LIMITATIONS

Our constructions establish that relaxing each structural property is sufficient to recover the corresponding behavioral dimension (Propositions 1–3); we do not establish that these are the only sufficient relaxations, and leave full necessity to future work.

Finite-precision score representations generically produce large tie orbits (Proposition 4): whenever a score is computed with a fixed, finite number of representable levels, the number of distinct score orderings sharing a given metric value grows exponentially in n. A practical consequence of Theorem 3 is that such collapse families are likely large and common in real-world evaluation, not rare edge cases, though we do not quantify actual tie-orbit sizes for deployed systems beyond the COMPAS reproduction, leaving that empirical characterization to future work.

The Collapse Audit gives an empirical lower bound on collapse-family spread, not an exact characterization; a zero-spread result establishes only that the sampled perturbations did not reveal collapse, not that no collapse family exists. Low resampling budget is a benign failure mode that resolves with more rounds, but a perturbation distribution restricted to a narrow neighborhood can underestimate true spread by up to $50\times$ even at large budget, and does not resolve with compute alone (Section 4.2).

The framework also requires knowing the behavioral dimension ϕ at stake in advance; practitioners may not know which behaviors matter until deployment, or may face several competing concerns at once. Automated discovery of which ϕ values are deployment-critical would strengthen practical applicability but remains open.

Our real-world evidence is one fully-reproduced instance, COMPAS; the anomaly-detection and language-modeling settings rely on supporting citations rather than an independently reproduced deployed system. Theorem 3's exact count applies only to the ordered score sequence and its injective functions; detection latency and subgroup disparity remain covered only by the asymptotic bounds of Theorems 1 and 2. Finally, the framework diagnoses when a metric cannot see a behavior; it does not itself propose replacement metrics, and selecting among existing remedies for a specific deployment remains a domain-specific judgment outside our scope. Broader societal implications are discussed in Appendix H.

7 CONCLUSION

Scalar summaries make comparison feel simple: one number, one ranking, one decision. We have shown this simplicity has a precise, provable cost: any metric built from pointwise scoring, permutation invariance, and scalar aggregation is structurally incapable of representing behavioral properties outside a characterizable information class, regardless of the scoring rule chosen. The impossibility is quantitative: collapse families grow as $\Theta(n)$ or faster and persist at the maximum attainable score, where practitioners are most confident in the result (Theorem 2), and no choice of scoring rule can fix what the evaluation representation discards (Corollary 1). For the ordered score sequence itself, the collapse-family size is not a bound but an exact orbit count (Theorem 3).

It is also not merely constructible: the same signature, near-identical aggregate performance concealing divergent subgroup behavior, appears in COMPAS, a real deployed system whose public controversy was, in effect, an argument over incompatible projections of one collapsed number, independently reproduced from public data.

The response we offer is not a new metric, but a discipline for using the ones we have: before a scalar score is treated as evidence for a behavioral claim, evaluation-time information should be shown sufficient for that claim, analytically via Propositions 1–3, or empirically via the Collapse Audit, run directly on scores already computed, with no retraining or new benchmark required. Every evaluation pipeline can be made to answer this question before, rather than after, a model is deployed on the strength of a number that was never positioned to tell the whole story: does the evaluation retain enough information to see the behavior it is being asked to certify?

AI USE STATEMENT

Generative AI tools were used to assist with hypothesis refinement, formalization of the structural properties and theorem statements, drafting and editing of the manuscript, implementation of the experimental code and the Collapse Audit procedure, and interpretation of results, including the independent reproduction of the COMPAS statistics. All AI-assisted material was reviewed and independently verified by the authors, including manual verification of every theorem and proof, re-execution of all reported experiments, and independent cross-checking of the reproduced real-world statistics against the original published sources. The authors take full responsibility for the final content of the work, including its text, mathematical claims, experimental results, analyses, and conclusions.

REPRODUCIBILITY STATEMENT

We provide an anonymous source-code repository containing the complete implementation, notebooks, and scripts used for all experiments, figures, and tables reported in this paper: `https://anonymous.4open.science/r/eval_collapse-CB6B/README.md`. The formal definitions and theorem statements are given in Section 3, with complete, expanded proofs in Appendix A; the three explicit constructions (AUROC/latency, accuracy/subgroup disparity, perplexity/trajectory) are derived in Section 3 and Appendix B; the Collapse Audit procedure is specified in Algorithm 1 (Section 4) with implementation details in Appendix E; robustness checks across continuous score distributions and random seeds are documented in Appendix C; and the independent reproduction of the COMPAS accuracy/FPR/FNR statistics, including the exact public data source and filtering criteria used, is documented in Appendix D. The exact orbit-counting result (Theorem 3) is fully proved in Appendix A.5, with extended scope discussion in Appendix A.6, its perplexity instantiation in Appendix A.7, and the bounded-precision pigeonhole bound in Appendix A.8; the Collapse Audit's false-negative failure-mode ablation is in Appendix F, and broader societal impact is discussed in Appendix H. The repository includes standalone Python scripts (`code/audit_core.py`, `code/compas_reproduction.py`) that reproduce every reported number without retraining any model, and a self-contained notebook (`notebooks/`) that reproduces all synthetic constructions, real-data validations (NAB, UCI Adult, GPT-2), and figures end to end.

REFERENCES

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, May 2016.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pp. 77–91. PMLR, 2018.

Mayee F. Chen, Michael Y. Hu, Nicholas Lourie, Kyunghyun Cho, and Christopher Ré. Aioli: A unified optimization framework for language model data mixing. *arXiv preprint arXiv:2411.05735*, 2024.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

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 (ICML)*, volume 70, pp. 1321–1330. PMLR, 2017.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Alexander Lavin and Subutai Ahmad. Evaluating real-time anomaly detection algorithms – the Numenta anomaly benchmark. In *Proceedings of the 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pp. 38–44. IEEE, 2015.

Andrew Wong, Erkin Otles, John P. Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8):1065–1070, 2021. doi: 10.1001/jamainternmed.2021.2626.

A Full Proofs

This appendix gives complete, expanded proofs and derivations for every theorem, corollary, and remark stated in Section 3, including the material moved here in full from the main text to respect the page limit.

A.1 Proof of Lemma 1 (Information-Theoretic Impossibility)

Proof. By hypothesis, ϕ is not determined by $\mathcal{E}$: there is no function h with $\phi(f, D) = h(\mathcal{E}(f, D))$ for all (f, D). Equivalently, there exist configurations (f, D) and (f', D') with

$$\mathcal{E}(f, D) = \mathcal{E}(f', D') \quad \text{but} \quad \phi(f, D) \neq \phi(f', D'),$$

since if no such pair existed, the map $\mathcal{E}(f, D) \mapsto \phi(f, D)$ would be well-defined (single-valued) on the range of $\mathcal{E}$, giving exactly the determining function h whose non-existence we assumed.

Now fix any metric $\mathcal{M} = G \circ \mathcal{E}$ for arbitrary $G : \mathcal{R} \to \mathbb{R}$. Since $\mathcal{E}(f, D) = \mathcal{E}(f', D')$, applying the same function G to both sides preserves the equality:

$$\mathcal{M}(f, D) = G(\mathcal{E}(f, D)) = G(\mathcal{E}(f', D')) = \mathcal{M}(f', D').$$

This holds for *every* choice of G, since the argument only used that G is a function (single-valued), not any property specific to a particular scoring rule. Hence no choice of G can separate (f, D) from (f', D') once $\mathcal{E}$ has merged them. □

Remark 3. *This proof makes precise the sense in which the impossibility is* structural *rather than a weakness of any specific metric: the argument never uses the functional form of G, only that $\mathcal{M}$ factors through $\mathcal{E}$. Any metric that is a function of $\mathcal{E}$ alone, which by definition includes accuracy, AUROC, F1, and perplexity once their respective $\mathcal{E}$ is fixed, inherits this limitation automatically.*

A.2 Proof of Theorem 1 (Linear Collapse), Expanded

Proof. Fix f, $D = \{(x_i, y_i)\}_{i=1}^n$, and let $s_i = s(f(x_i), y_i)$ denote the per-instance score sequence, so that by pointwise evaluation, $\mathcal{E}(f, D) = H(s_1, \ldots, s_n)$.

For a permutation σ of $[n]$, write $\sigma D = \{(x_{\sigma(i)}, y_{\sigma(i)})\}_{i=1}^n$ for the dataset with instances reordered according to σ. Since s is computed pointwise (Definition 1), the score sequence for σD is exactly $(s_{\sigma(1)}, \ldots, s_{\sigma(n)})$, a reordering of the original sequence, with no new values introduced and none removed. By permutation invariance (Definition 2),

$$H(s_1, \ldots, s_n) = H(s_{\sigma(1)}, \ldots, s_{\sigma(n)}),$$

so $\mathcal{E}(f, \sigma D) = \mathcal{E}(f, D)$ for *every* permutation σ, not merely the k_n permutations $\sigma_1, \ldots, \sigma_{k_n}$ posited in the theorem statement. Restricting to these k_n permutations and applying G,

$$\mathcal{M}(f, \sigma_j D) = G(\mathcal{E}(f, \sigma_j D)) = G(\mathcal{E}(f, D)) = \mathcal{M}(f, D) = r \qquad \text{for all } j = 1, \ldots, k_n.$$

Thus all k_n configurations $\{(f, \sigma_j D)\}_{j=1}^{k_n}$ lie in the collapse family $\mathcal{C}_r$. By hypothesis, $\phi(f, \sigma_i D) \neq \phi(f, \sigma_j D)$ for all $i \neq j$, so these $k_n = \Theta(n)$ configurations are pairwise behaviorally distinct, giving $|\mathcal{C}_r| \geq \Theta(n)$. □

A.3 Proof of Theorem 2 (Collapse at Maximum Score), Expanded

Proof. Suppose the linear collapse family $\{(f, \sigma_j D)\}_{j=1}^{k_n}$ from Theorem 1 is optimal-score: some $j^* \in \{1, \ldots, k_n\}$ satisfies $\mathcal{M}(f, \sigma_{j^*} D) = r^* := \max_{f,D} \mathcal{M}(f, D)$.

As shown in the proof of Theorem 1, permutation invariance gives $\mathcal{E}(f, \sigma_j D) = \mathcal{E}(f, \sigma_{j^*} D)$ for every $j \in \{1, \ldots, k_n\}$, since both are permutations of the same underlying dataset D and hence lie in the same orbit under the pointwise-then-permute construction. Applying G,

$$\mathcal{M}(f, \sigma_j D) = G(\mathcal{E}(f, \sigma_j D)) = G(\mathcal{E}(f, \sigma_{j^*} D)) = \mathcal{M}(f, \sigma_{j^*} D) = r^*$$

for all j. Since r^* is by definition the maximum attainable value of $\mathcal{M}$, and every configuration in the orbit attains exactly r^*, the collapse family at the maximum score $\mathcal{C}_{r^*}$ contains all $k_n = \Theta(n)$ configurations $\{(f, \sigma_j D)\}$. These remain pairwise distinct in ϕ by the hypothesis inherited from Theorem 1. □

A.4 Proof of Corollary 1 (Optimality is Not Self-Correcting)

Proof. Let $\{(f, \sigma_j D)\}_{j=1}^{k_n}$ be an optimal-score collapse family as in Theorem 2, so every configuration in the orbit satisfies $\mathcal{E}(f, \sigma_j D) = \mathcal{E}(f, \sigma_{j^*} D) =: e^*$ for a fixed representation value e^*. For *any* choice of scoring rule $G' : \mathcal{R} \to \mathbb{R}$ (not necessarily the original G), the induced metric $\mathcal{M}' = G' \circ \mathcal{E}$ satisfies

$$\mathcal{M}'(f, \sigma_j D) = G'(\mathcal{E}(f, \sigma_j D)) = G'(e^*)$$

for every j, since all k_n configurations share the identical representation value e^*; this is a fact about $\mathcal{E}$, independent of which G' is chosen downstream. Hence every G' maps the entire orbit to the single value $G'(e^*)$, and no choice of G' can assign different scores to configurations that $\mathcal{E}$ has already merged. Separating them requires changing $\mathcal{E}$ itself, e.g. via one of the relaxations in Propositions 1–3. □

A.5 Full Statement and Proof of Theorem 3 (Exact Orbit Collapse)

We restate the score-sequence orbit definition and give the theorem's complete proof, including parts compressed to a sketch in the main text.

Definition 5 (Score-Sequence Orbit)**.** *Let $s = (s_1, \ldots, s_n)$ be the per-instance score sequence for (f, D), with $k \leq n$ distinct values occurring with multiplicities $n_1, \ldots, n_k$ ($\sum_j n_j = n$). The* score-sequence orbit *is*

$$\mathrm{Orb}(s) \;=\; \big\{\, (s_{\sigma(1)}, \ldots, s_{\sigma(n)}) \;:\; \sigma \in S_n \,\big\},$$

the set of distinct sequences *obtained by permuting the coordinates of s: a set of sequences, not a multiset indexed by σ, so sequences produced by different permutations that happen to coincide are counted once.*

Proof of Theorem 3. **(a) Orbit size.** The symmetric group S_n acts on the set of coordinate sequences by $\sigma \cdot (s_1, \ldots, s_n) = (s_{\sigma(1)}, \ldots, s_{\sigma(n)})$. The stabilizer of s under this action is

$$\mathrm{Stab}(s) = \{\sigma \in S_n : (s_{\sigma(1)}, \ldots, s_{\sigma(n)}) = (s_1, \ldots, s_n)\}.$$

A permutation σ fixes the sequence s pointwise if and only if it permutes indices only *within* each tie class: moving an index from one tie class to another would change the value at that coordinate, since scores across classes differ. Hence $\mathrm{Stab}(s) \cong S_{n_1} \times S_{n_2} \times \cdots \times S_{n_k}$, the direct product of symmetric groups on each tie class, of order $\prod_{j=1}^{k} n_j!$. By the orbit–stabilizer theorem, $|\mathrm{Orb}(s)| = |S_n| / |\mathrm{Stab}(s)| = n! / \prod_j n_j!$.

(b) Metric invariance over the full group. By pointwise evaluation, permuting the instances of D by any $\sigma \in S_n$ permutes the coordinates of the score sequence identically: the score sequence of $\sigma \cdot D$ is $(s_{\sigma(1)}, \ldots, s_{\sigma(n)})$, since $s_{\sigma(i)} = s(f(x_{\sigma(i)}), y_{\sigma(i)})$ depends only on instance $\sigma(i)$'s own output and label. By permutation invariance of H,

$$\mathcal{E}(f, \sigma \cdot D) = H(s_{\sigma(1)}, \ldots, s_{\sigma(n)}) = H(s_1, \ldots, s_n) = \mathcal{E}(f, D)$$

for *every* $\sigma \in S_n$. This is exactly the argument of Theorem 1's proof (Appendix A), but applied without restricting to a constructed subfamily $\sigma_1, \ldots, \sigma_{k_n}$; it holds for the entire group. Applying any scoring rule G preserves this equality, so $\mathcal{M}(f, \sigma \cdot D) = \mathcal{M}(f, D)$ for every $\sigma \in S_n$.

(c) Exact count under ϕ_{ord}. Define $\phi_{\mathrm{ord}}(f, D) \coloneqq (s_1, \ldots, s_n)$, the ordered score sequence. For $\sigma, \sigma' \in S_n$,

$$\phi_{\mathrm{ord}}(f, \sigma{\cdot}D) = \phi_{\mathrm{ord}}(f, \sigma'{\cdot}D) \iff (s_{\sigma(1)}, \ldots, s_{\sigma(n)}) = (s_{\sigma'(1)}, \ldots, s_{\sigma'(n)}) \iff \sigma^{-1}\sigma' \in \mathrm{Stab}(s),$$

i.e. σ and σ' lie in the same left coset of $\mathrm{Stab}(s)$ in S_n. The number of distinct left cosets is $[S_n : \mathrm{Stab}(s)] = n!/\prod_j n_j! = |\mathrm{Orb}(s)|$ by part (a), and each coset maps to exactly one distinct value of ϕ_{ord}, so the $n!$ configurations $\{(f, \sigma \cdot D)\}_{\sigma \in S_n}$ take exactly $|\mathrm{Orb}(s)|$ pairwise distinct values under ϕ_{ord}. By part (b), all $n!$ configurations share the single metric value $r = \mathcal{M}(f, D)$. Hence the collapse family $\mathcal{C}_r$ contains at least these $n!$ configurations, of which exactly $|\mathrm{Orb}(s)| = n!/\prod_j n_j!$ are pairwise behaviorally distinct under ϕ_{ord}: neither more, since ϕ_{ord} can take no more values than $|\mathrm{Orb}(s)|$ by definition of the orbit, nor fewer, since every coset is realized by some $\sigma \in S_n$. □

A.6 Extended Scope Discussion for Theorem 3

Theorem 3 is exact for the behavioral dimension ϕ_{ord} and, more generally, for any $\psi = g \circ \phi_{\mathrm{ord}}$ with g injective: since ϕ_{ord} already partitions the $n!$ configurations into exactly $|\mathrm{Orb}(s)|$ classes, composing with an injective g preserves this partition exactly, so ψ also takes exactly $|\mathrm{Orb}(s)|$ distinct values across the same configurations.

This scope does *not* automatically extend to *many-to-one* functions of ϕ_{ord}, which is the case for several behavioral dimensions used elsewhere in this paper:

- **Detection latency**, $\phi_{\text{latency}} = \min\{t : s_t > \tau\} - t_0$, depends only on the position of the *first* above-threshold score, not on the full ordering. Many distinct permutations of the tail (after the first crossing) share the same latency value, so ϕ_{latency} is a strict many-to-one function of ϕ_{ord}, and the exact orbit count of Theorem 3 overcounts the true number of behaviorally distinct latency values. The AUROC/latency construction (Appendix B) is accordingly covered only by the asymptotic bound of Theorem 2, which directly counts distinct latency values via the constructed subfamily $\sigma_1, \ldots, \sigma_{k_n}$ rather than via the full orbit.
- **Subgroup disparity**, $\Delta_j = |k - 2j|/m$, depends only on the count j of errors allocated to group 1, not on which specific instances are misclassified. Many distinct assignments of individual errors to instances share the same count j, so Δ_j is similarly many-to-one in the underlying score/label assignment. The accuracy/disparity construction is covered only by the asymptotic bound of Theorem 1.
- **First-passage time and run-length statistics** of a score sequence are likewise many-to-one functions of the full ordering: many distinct sequences share the same first-passage time or the same longest run of low-confidence scores. An exact collapse rate for such statistics would require a separate injectivity (or exact fiber-counting) argument specific to that statistic, which we do not attempt in this paper.

Establishing exact (rather than asymptotic) collapse rates for these coarser statistics is a natural direction for future work, following the same orbit-counting strategy but requiring the size of each *fiber* of the many-to-one map (i.e. how many orderings map to each latency, disparity, or run-length value) rather than the stabilizer of the full sequence.

A.7 Perplexity as an Exact Instance of Theorem 3

The perplexity/trajectory construction of Section 3.7.4 sets $\phi_{\text{traj}}(f, D) = (\ell_{\sigma(1)}, \ldots, \ell_{\sigma(T)})$, the ordered sequence of per-token log-probabilities. By definition, $\phi_{\text{traj}} = \phi_{\mathrm{ord}}$ exactly, with no coarsening. Taking $n = T$ and tie multiplicities $n_1 = m$ (tokens at log-probability ℓ_1), $n_2 = T - m$ (tokens at ℓ_2), Theorem 3 gives collapse-family size exactly

$$|\mathrm{Orb}(s)| = \frac{T!}{m!\,(T-m)!} = \binom{T}{m},$$

matching the combinatorial count derived directly in Appendix B. At $m = T/2$, Stirling's approximation gives $\binom{T}{T/2} = \Theta(2^T/\sqrt{T})$. This confirms that the $\binom{T}{m}$ collapse-family size for perplexity is not an independent combinatorial fact about this particular construction, but a direct special case ($k = 2$ tie classes) of the general orbit–stabilizer formula of Theorem 3.

A.8 Proof of the Pigeonhole Bound (Ties Are Common for Bounded-Precision Scores)

Proposition 4 (Generic Exponential Tie Orbits under Bounded Precision)**.** *Suppose the score function s takes at most L distinct values on a dataset of size n, for some fixed $L \geq 2$. Then the tie orbit $\mathrm{Orb}(s)$ has size*

$$|\mathrm{Orb}(s)| \;\geq\; \binom{n}{\lceil n/L \rceil} \;=\; \Theta\!\left(\frac{2^{nH(1/L)}}{\sqrt{n}}\right) \quad \text{as } n \to \infty \text{ for fixed } L,$$

where $H(p) = -p\log_2 p - (1-p)\log_2(1-p)$ is the binary entropy function, regardless of how the n instances are distributed among the L score levels.

Proof. Let the score sequence s partition the n instances into $k \leq L$ tie classes of sizes $n_1, \ldots, n_k$ (with $\sum_j n_j = n$), so that by Theorem 3(a), $|\mathrm{Orb}(s)| = n!/\prod_{j=1}^{k} n_j!$.

By the pigeonhole principle, some class has multiplicity $n_{\max} \geq \lceil n/L \rceil$. Relabel so that $n_1 = n_{\max}$, and let $m = n - n_{\max} = n_2 + \cdots + n_k$ denote the combined size of the remaining classes.

The quantity $m!/(n_2! \cdots n_k!)$ is the multinomial coefficient counting arrangements of m items into groups of sizes $n_2, \ldots, n_k$, hence a positive integer, so

$$n_2!\, n_3! \cdots n_k! \;\leq\; m! \;=\; (n - n_{\max})!.$$

Multiplying both sides by $n_{\max}!$ gives

$$\prod_{j=1}^{k} n_j! \;=\; n_{\max}! \cdot \big(n_2! \cdots n_k!\big) \;\leq\; n_{\max}!\,(n - n_{\max})!.$$

Taking reciprocals and multiplying by $n!$,

$$|\mathrm{Orb}(s)| \;=\; \frac{n!}{\prod_{j=1}^{k} n_j!} \;\geq\; \frac{n!}{n_{\max}!\,(n - n_{\max})!} \;=\; \binom{n}{n_{\max}}.$$

Since $\binom{n}{\cdot}$ is increasing on $[0, n/2]$ and $n_{\max} \geq \lceil n/L \rceil \leq n/2$ for $L \geq 2$, this is in turn at least $\binom{n}{\lceil n/L \rceil}$. By Stirling's approximation, for $n_{\max} = n/L$ a fixed fraction of n,

$$\binom{n}{n/L} = \Theta\!\left(\frac{2^{nH(1/L)}}{\sqrt{n}}\right).$$

This bound depends only on L and n, not on the specific distribution of the n instances among the (at most) L score levels, establishing the claim. □

Remark 4. *Proposition 4 formalizes the informal claim that large tie orbits are the* generic*, not exceptional, regime for any evaluation with a fixed, finite number of representable score levels, including binary classifiers ($L = 2$), any score rounded or quantized to L representable values, and, in the limit, any score computed in finite floating-point precision once n exceeds the number of distinct representable values in the relevant range. The bound is a worst-case-favorable (i.e. guaranteed lower bound), not a claim about the* typical *orbit size for an arbitrary or adversarial score distribution, which could be larger still.*

B Full Derivations for the Explicit Constructions

B.1 AUROC and Detection Latency

Recall the setup: T time steps, anomaly onset at t_0, score bands $s_{\text{normal}} < s_{\text{pre-alert}} < \tau < s_{\text{alert}}$. For alert position $t \in \{t_0, \ldots, t_0 + (T - t_0)/2\}$, all t_0 normal instances receive scores below s_{normal}, and all $T - t_0$ anomalous instances receive scores above s_{normal} (either in the pre-alert band or the alert band). Since every normal score lies strictly below every anomalous score regardless of t, the ROC curve passes through the point $(0, 1)$ for every threshold between s_{normal} and $\min(\text{anomalous scores})$, giving AUROC exactly 1.000 for all $(T - t_0)/2 + 1 = \Theta(T)$ choices of t. Detection latency, defined as $\min\{i \geq t_0 : s_i > \tau\} - t_0$, equals exactly $t - t_0$ by construction, so as t ranges over its $\Theta(T)$ values, latency ranges over $\{0, 1, \ldots, (T - t_0)/2\}$, a set of $\Theta(T)$ pairwise distinct values.

B.2 Accuracy and Subgroup Disparity

With two groups of size m and $k = \Theta(m)$ total errors split as j in group 1 and $k - j$ in group 2 for $j \in \{0, \ldots, k\}$: total errors are $j + (k - j) = k$ regardless of j, so overall accuracy over $2m$ instances is $\mathrm{Acc} = 1 - k/(2m)$, constant in j. Subgroup accuracies are $\mathrm{Acc}_1 = 1 - j/m$ and $\mathrm{Acc}_2 = 1 - (k - j)/m$, giving disparity

$$\Delta_j = |\,\mathrm{Acc}_1 - \mathrm{Acc}_2\,| = \left|\frac{k-j}{m} - \frac{j}{m}\right| = \frac{|k-2j|}{m}.$$

This is minimized at $j = k/2$ (giving $\Delta = 0$) and maximized at $j \in \{0, k\}$ (giving $\Delta = k/m$), so as j ranges over its $k + 1 = \Theta(m)$ integer values, Δ_j takes $\Theta(m)$ distinct values in $[0, k/m]$ (exactly $\lfloor k/2 \rfloor + 1$ distinct values by symmetry of $|k - 2j|$ around $j = k/2$).

The perplexity/trajectory construction's derivation is given separately in Appendix A.7, since it is a direct instance of the exact orbit-counting theorem (Theorem 3) rather than an asymptotic bound.

C Robustness Checks: Full Details

C.1 Continuous-Score Robustness

To confirm the collapse phenomenon is not an artifact of the hand-placed score bands used for exposition in Section 3, we repeat each construction with scores drawn from continuous distributions:

- **Anomaly detection:** normal scores $\sim$ Beta$(2, 8)$ (concentrated near 0), pre-alert anomaly scores $\sim$ Beta$(3, 5)$, alert-band scores $\sim$ Beta$(8, 2)$ (concentrated near 1), rescaled to preserve the band ordering $s_{\text{normal}} < s_{\text{pre-alert}} < \tau < s_{\text{alert}}$ in expectation and with high probability.
- **Subgroup evaluation:** identical to the closed-form construction (errors are discrete Bernoulli outcomes, so no continuous relaxation is applicable; robustness here is instead checked via random seed, see below).
- **Language modeling:** log-probabilities drawn from $\mathcal{N}(\ell_1, 0.05)$ and $\mathcal{N}(\ell_2, 0.05)$ respectively, rather than fixed constants.

In all cases, the collapse signature is unchanged to within floating-point precision: AUROC remains exactly 1.000 across all sampled latencies (up to the 10^{-10} tolerance of the underlying ROC computation), and perplexity remains invariant across all sampled trajectories to the same tolerance.

C.2 Seed Robustness

We repeat the subgroup-disparity construction across 20 independent random seeds and the trajectory construction across 10 independent seeds. In both cases, the resulting disparity-vs-j and perplexity-vs-seed curves overlap to within floating-point precision across all seeds (visually indistinguishable; see the released notebook for the overlaid plots), confirming the collapse is a structural consequence of the metric's definition rather than an artifact of a specific random draw.

D COMPAS Reproduction: Methodology

We use ProPublica's publicly released `compas-scores-two-years.csv` dataset and apply their published filtering criteria exactly as documented in their methodology write-up "How We Analyzed the COMPAS Recidivism Algorithm": we retain only records with `days_b_screening_arrest` between -30 and 30 inclusive (to exclude cases where the COMPAS screening date is not closely aligned with the arrest date), `is_recid` $\neq -1$ (excluding cases with missing recidivism outcome), `c_charge_degree` $\neq$ `'O'` (excluding ordinary traffic offenses, which do not carry jail time and are not the target population), and `score_text` $\neq$ `'N/A'` (excluding missing risk scores). This yields $n = 6{,}172$ records after filtering, of which 2,103 are recorded as "Caucasian" and 3,175 as "African-American" in the dataset's `race` field (the two groups compared in Table 2; other race categories present in the data are omitted from this specific comparison for direct comparability with the original analysis).

We define a positive prediction as a COMPAS `score_text` of "Medium" or "High" risk (matching the original analysis's operationalization of the risk score as informing a binary high-risk / low-risk decision), and the ground-truth positive as `two_year_recid` $= 1$ (any recorded recidivism within two years). Accuracy, false-positive rate, and false-negative rate are computed via the standard confusion-matrix definitions on each racial subgroup independently. Full reproduction code is provided in `code/compas_reproduction.py` in the released repository, requiring only the public CSV as input, with no proprietary or restricted-access data.

E Collapse Audit: Implementation Details

Algorithm 1 is implemented generically in `code/audit_core.py`: the core routine `collapse_audit(observed_score, metric_fn, phi_fn, perturbation_fn, B)` accepts arbitrary callables for the metric, the behavioral dimension, and the perturbation distribution, and returns the empirical spread, standard deviation, and count of metric-invisible perturbations found within the budget B. For the results in Table 3, we use $B = 3000$ and the following per-setting perturbation distributions: for AUROC/latency, uniform resampling of the alert position t over its valid range; for accuracy/disparity, uniform resampling of the error allocation $j \in \{0, \ldots, k\}$; for perplexity/trajectory, uniform random shuffling of the fixed log-probability multiset. All three audits found $3000/3000$ perturbations metric-invisible (i.e. every sampled perturbation left the observed score exactly unchanged, as guaranteed analytically by Theorems 1–2 for these constructions), so the reported empirical spreads in Table 3 use the full sample.

F Collapse Audit: Failure Mode Ablation

We report the full ablation supporting Section 4.2, run on both the subgroup-disparity and AUROC/latency constructions.

F.1 Budget Ablation

For the subgroup-disparity construction ($m = 500$, $k = 100$, true spread $k/m = 0.20$), we sweep $B \in \{1, 3, 5, 10, 25, 50, 100, 500, 1000, 3000\}$ using the full-range perturbation $j \sim \text{Uniform}\{0, \ldots, k\}$. Empirical spread reaches 50% of the true value by $B = 3$, 98% by $B = 25$, and 100% (to floating-point precision) by $B = 50$. This confirms that budget-driven false negatives are transient and resolve quickly once B exceeds a small constant multiple of the collapse family's diversity.

F.2 Perturbation-Locality Ablation

Fixing $B = 3000$, we restrict the perturbation to a neighborhood of radius r around the midpoint allocation $j_0 = k/2$ (the zero-disparity configuration), i.e. $j \sim \text{Uniform}\{\max(0, j_0 - r), \ldots, \min(k, j_0 + r)\}$, rather than the full range $\{0, \ldots, k\}$. Table 4 reports the resulting spread as a fraction of the true value.

Table 4: Perturbation-locality ablation, subgroup-disparity construction, $B = 3000$.

Radius r	Empirical spread	Fraction of true spread (0.20)
1	0.0040	2%
2	0.0080	4%
5	0.0200	10%
10	0.0400	20%
25	0.1000	50%
50 (full range)	0.2000	100%

We confirm the same qualitative pattern on the AUROC/latency construction ($T = 200$, $t_0 = 100$, true theoretical spread 50 steps): restricting the perturbation to a neighborhood of radius r around the midpoint latency yields empirical spreads of 4%, 8%, 20%, and 40% of the true value at $r \in \{1, 2, 5, 10\}$ respectively, reaching 100% only once r spans the full latency range.

F.3 INTERPRETATION

Both ablations use the same underlying construction and the same true collapse family; only the audit's own settings change. The budget ablation shows the audit converges reliably given a sufficiently wide perturbation family. The locality ablation shows that no amount of additional budget compensates for a perturbation distribution that does not reach the relevant region of metric-invisible configurations. This is a qualitatively different, non-resolvable failure mode, and is the primary caveat we attach to the audit's negative results (Section 4.2 and Section 6).

G PRACTICAL RECOMMENDATIONS TABLE

Table 5 expands on Section 5's practical protocol, giving the specific remedy and marginal cost for each behavioral concern.

Table 5: Structural relaxations: which to apply for each behavioral concern.

Concern	Relax	Remedy	Cost
Detection delay	Perm. inv.	Retain order; report delay-weighted scores (e.g., NAB)	Free
Subgroup harm	Pointwise	Retain labels; report per-group rates	Free
Confidence shape	Scalar agg.	Retain score sequence; report calibration curves	Small
Unknown ϕ or legacy audit	—	Run Collapse Audit (Alg. 1)	$O(n)$

H BROADER IMPACT

Evaluation metrics do not just measure models; they shape what gets built, deployed, and trusted. When a scalar score is treated as a complete certificate of behavior, the collapse phenomenon we formalize means that decisions (clinical deployment, sentencing-adjacent risk tools, model leaderboards) can be made on the strength of a number that was never capable of ruling out the behavior that mattered. This is not a hypothetical risk: the COMPAS case study in Section 3.7.3 shows it has already happened in a system with direct consequences for real people. As AI systems are increasingly evaluated and gated on aggregate benchmark scores (including for safety-relevant properties such as harmful-output rates, jailbreak resistance, or bias), the same structural blindness applies: a model can score identically to a safer or more harmful counterpart on a benchmark that is pointwise, permutation-invariant, and scalar, with the difference invisible until deployment surfaces it. The Collapse Audit (Section 4) is offered as a low-cost mitigation: it requires no new benchmark or retraining, only a check, on data already collected, of whether a reported score is actually evidence for the claim being made from it. We see the primary societal value of this work not in any single result but in the discipline it proposes: treating "does this evaluation retain enough information?" as a routine question before a score is used to justify a deployment decision, rather than an afterthought raised only once something has already gone wrong.