FullSongEval: Auditing Text–Label Associations Before Music Evaluation

Anonymous authors
Paper under double-blind review

Abstract

Public music datasets pair text with labels such as genre. Before using these labels in evaluation, researchers may ask whether the text predicts them. Yet a scorer can remain above label prevalence after complete label sets are reassigned among records. FullSongEval audits a candidate rule that accepts text–label pairings when a paired fold interval for a scorer's average-precision margin over prevalence has a positive lower endpoint. We test original-pair acceptance and scorer sensitivity to an injected label-linked cue. We then refit the scorer after two fixed label-set reassignments per dataset and ask whether the rule rejects them. On MusicCaps and MTG-Jamendo, the scorer detects the cue, but the rule accepts both reassignments. On the same MTG tracks, audio-feature genre classifiers trained with original tags predict independently collected consensus genres more accurately than classifiers trained with reassigned tags: mean average precision is 0.7927 versus 0.4365 and 0.4304. A diagnostic finite-sample random-label reference accounts for tied scores. In a separate 8,000-record Free Music Archive check fixed before source inspection, this reference separates original pairings from two reassignment controls, while the prevalence rule accepts all three. Exceeding prevalence is insufficient to approve these text–label pairings for later music evaluation.

1 Introduction

Public captions and catalog tags offer inexpensive supervision for music evaluation (Agostinelli et al., 2023; Bogdanov et al., 2019). A benchmark might use them to test whether a model recognizes genre, mood, or instrumentation. Before using these labels to evaluate a music system, researchers can check whether text about a recording predicts its attached labels. This check does not measure audio quality, listener preference, or the quality of a complete music system. Nor does it establish what inputs current song-evaluation systems use. FullSongEval[1] tests one rule for deciding whether this text–label association is credible enough to use in later evaluation.

A high text-prediction score can be misleading. The scorer may exploit frequent labels or wording shortcuts, and a rule that accepts such a score may still accept data after text and labels have been mismatched. If the accepted data are then used to select evaluation examples or compare systems, the error can affect those later decisions.

The check has two parts. First, after direct label words are removed, the scorer must respond when we add a phrase or codeword linked to a label. This tests *scorer sensitivity*. Second, the rule must reject data after we reassign label sets among records, disrupting the original text–label pairings.

Precision–recall analyses sometimes use label prevalence as a reference for average precision (Saito & Rehmsmeier, 2015; Boyd et al., 2013). We test a candidate rule based on that comparison: it accepts the text and labels when the scorer exceeds a constant predictor that uses only label frequency. We call this the *prevalence reference*; under our grouped-tie calculation it equals held-out label prevalence, even though its constant score is set from the training fold. This exact acceptance rule is not established as a field standard. It tests the narrower inference that a score above prevalence suffices

[1]Anonymous code repository: `https://anonymous.4open.science/r/full-song-eval-anonymous-B204/README.md`. The supplementary archive, `fullsongeval-anonymous-source.zip`, contains the complete runnable reproduction package and lightweight inputs.

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.

to approve an association. We ask whether the scorer detects a known cue, whether the rule accepts original pairings but rejects reassigned labels, and whether a finite-sample reference that accounts for score ties changes those decisions.

FullSongEval tests the rule on MusicCaps and MTG-Jamendo with known cues and two fixed complete-label-set reassignments per dataset. We compare the fitted scorer with finite-sample random-label expectations (Bestgen, 2015), accounting for ties. This diagnostic was developed after the primary audits but specified before the later FMA check.

Figure 1 shows the audit: accept original pairings, detect a known cue, and reject reassigned labels. The later random-label comparison does not replace the tested rule.

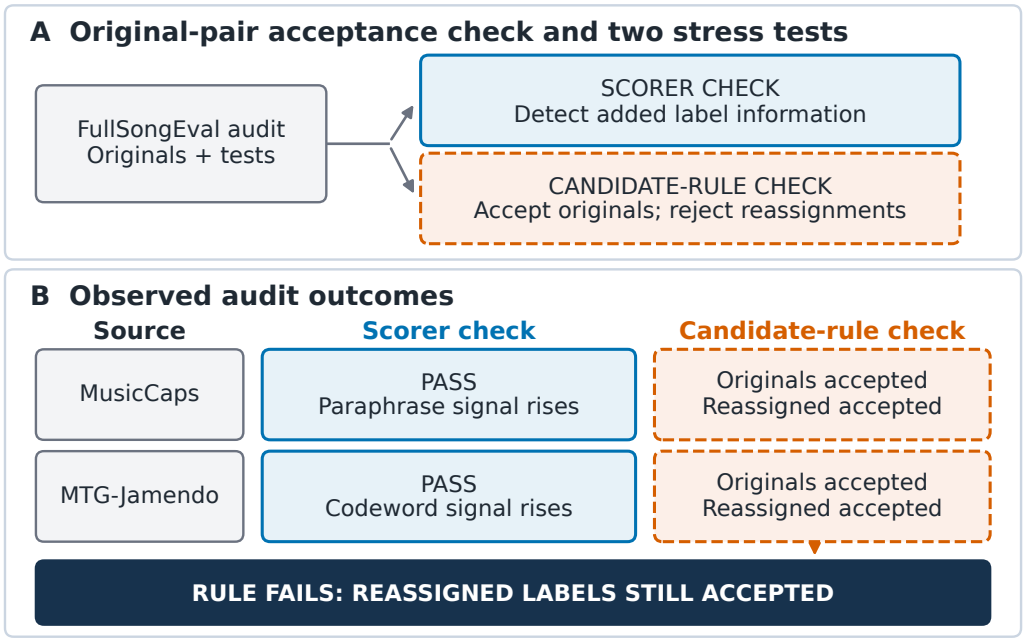

Figure 1: The audit requires the rule to accept original pairings, then tests the scorer and the rule separately. We add label-linked information without restoring exact label words, then cyclically reassign complete label sets using two fixed seeds. On both MusicCaps and MTG-Jamendo, the rule accepts the originals and the scorer responds to the added information, but the rule also accepts both reassignments. This failure motivates the later reference analysis in Equations 3 and 4.

The scorer detects added cues, but the rule accepts all four reassignment controls. On the same MTG tracks, reassigned tags weaken text and audio-feature prediction of independently collected consensus genres. A post-audit tie-aware reference withholds acceptance for these controls; new Song Describer captions and a separate FMA catalog add scoped checks (Appendix A.6).

Contributions. We make four contributions:

- We provide an audit that checks whether the scorer responds to known cues, the candidate rule accepts original pairings, and the rule withholds acceptance after label reassignment.
- We apply the known finite-sample expected AP for a random strict ranking and derive its counterpart for our grouped-tie scorer, showing why a constant-score prevalence reference is mismatched.
- In author-held-out MusicCaps and artist-held-out MTG folds, we test matched masking, injected cues, and fixed label-set reassignments.
- The rule accepts four reassignments whose MTG scorers and audio-feature classifiers predict consensus genres less accurately. A later tie-aware analysis withholds acceptance for these and the Song Describer and FMA controls, but remains diagnostic.

2 Related Work

Music and song evaluation. MusicCaps paired short music excerpts with natural-language descriptions for text-conditioned generation (Agostinelli et al., 2023). Recent datasets and benchmarks extend evaluation toward song aesthetics, multi-aspect quality, and human preference (Yao et al., 2025; Wu et al., 2026; Huang et al., 2025). Audio-text models also support scalable semantic comparisons (Huang et al., 2022; Elizalde et al., 2023). MIR studies have documented noisy public music tags and their effects on audio-tagging evaluation (Choi et al., 2018), while reference-based music metrics can be sensitive to the composition of the reference collection (Gui et al., 2024). These resources broaden what can be measured. Our primary audit asks whether public text and tags pass basic validity checks before later evaluation. A separate audio-feature genre classifier tests one consequence of training on tags admitted by the audited rule; it is not proposed as a music-quality metric.

Weak supervision and benchmark validity. Weak-supervision systems depend on the quality of their label sources (Ratner et al., 2020; Zhang et al., 2021). Benchmark studies have also shown that label errors, leakage, and shortcut features can distort evaluation (Northcutt et al., 2021; Kaufman et al., 2012; Geirhos et al., 2020). Control tasks help distinguish information in a representation from artifacts of the probe or dataset (Hewitt & Liang, 2019). In weak-label NLP benchmarks, Shao (2026) shuffles evidence among items while keeping queries and labels fixed to test whether model accuracy depends on evidence identity. Our primary controls reassign music metadata label sets in training and held-out records, refit the text scorer, and test whether a specified average-precision cutoff accepts them. A separate null analysis holds scores fixed. Neither analysis shows that the text predictions depend on audio evidence or listener judgments.

Measurement and benchmark audits. Measurement work separates a construct from the proxy used to operationalize it, while benchmark audits ask whether a task supports the interpretation attached to its score (Jacobs & Wallach, 2021; Raji et al., 2021). Reviews of music-generation evaluation similarly distinguish computational scores from the broader qualities they are intended to represent (Yang & Lerch, 2020). We study a rule that decides whether text and labels can be used in later evaluation. We test it by adding a known cue and by reassigning labels, not by proposing a new music-quality metric.

Precision–recall calibration. Precision–recall summaries are especially informative under label imbalance, but their baselines and finite-sample behavior require care (Saito & Rehmsmeier, 2015; Boyd et al., 2013). Bestgen (2015) showed that the exact expected AP of a random strict ranking exceeds label prevalence in finite samples. A constant score produces AP equal to held-out prevalence under the grouped-tie calculation used here. Our audit applies this mismatch to a decision about whether to use the data. We also compute the random-label expectation for the scorer's observed pattern of tied scores. We selected this comparison after the original rule failed, so it is diagnostic.

3 Calibration Protocol

3.1 Candidate rule and audit requirements

Each metadata record contains text, a set of target labels, and a group identifier. For MusicCaps, the text is a caption and the group is its caption author. For MTG-Jamendo, the text combines track, artist, and album fields, and the group is the artist. Within each training fold, we select the 16 most frequent labels that occur at least 25 times, breaking count ties alphabetically. We evaluate held-out records with macro area under the precision–recall curve (macro AUPRC), which gives equal weight to each label.

The candidate rule compares the text scorer with the prevalence reference: a constant-score predictor whose score is set from training-label frequency, but whose grouped-tie average precision on held-out records equals held-out label prevalence. Let A_c be the held-out macro AUPRC for condition c and A_{freq} this reference. Across ten fixed folds, we compute a paired 95% interval for

$$\Delta_c = A_c - A_{\text{freq}} \tag{1}$$

The rule accepts condition c for further evaluation when this interval's lower endpoint is above zero. This establishes a margin over the reference on these folds. If the lower endpoint is not above zero,

the condition is simply not accepted; this does not show that the rule detected a mismatch. We also require acceptance of the original text–label pairings so that a rule that rejects every condition cannot pass the reassignment check automatically. Accepting the original pairings shows that the rule is not trivially rejecting every condition, but it does not show that the original metadata are valid for audio or listener evaluation.

FullSongEval applies the original-pair check and two distinct stress tests. The scorer-sensitivity test compares target-masked text with the same text after injection and passes when the lower endpoint of the interval for $\text{AUPRC}_{\text{injected}} - \text{AUPRC}_{\text{reference}}$ is above zero. This comparison verifies that the scorer can respond to a known cue; it is not part of the candidate rule. The reassigned-label test applies Equation 1 to reassigned labels and passes when the rule does not accept them. The audit passes only if the rule accepts the original pairings, the scorer responds to the cue, and the rule rejects both reassignment controls. The later tie-aware comparison is diagnostic and does not affect this audit decision.

3.2 Scorer and decision calculation

For each label and training fold, we fit a deterministic bag-of-tokens log-odds scorer. Text is lowercased and split with the regular expression `[a-z0-9]+`; one-character tokens and the artifact's fixed 23-word stoplist are removed. Each remaining token is counted at most once per record. Let n_ℓ and $n_{\neg\ell}$ be the numbers of positive and negative training records for label ℓ, and let $n_{t,\ell}$ and $n_{t,\neg\ell}$ be the corresponding record counts for token t. With $\alpha = 0.5$, the score for the unique-token set $V(x)$ is

$$s_\ell(x) = \log \frac{n_\ell + \alpha}{n_{\neg\ell} + \alpha} + \sum_{t \in V(x)} \log \frac{(n_{t,\ell} + \alpha)/(n_\ell + 2\alpha)}{(n_{t,\neg\ell} + \alpha)/(n_{\neg\ell} + 2\alpha)}. \tag{2}$$

We rank held-out records with s_ℓ, compute average precision after grouping tied scores, and average across labels that have at least one positive held-out record. The prevalence reference assigns every held-out record the same smoothed training prevalence, $(n_\ell + 0.5)/(N + 1)$, and uses the same grouped-tie average-precision calculation.

Within each fold, we compute average precision for every selected label with at least one positive held-out record and then average those values. MusicCaps contributes 5–13 evaluated labels per fold; MTG-Jamendo contributes all 16. Each reported score is the arithmetic mean of the ten fold-level macro AUPRC values. For each fold, we subtract the reference value from the condition value and report $\bar{\Delta} \pm t_{0.975,9} s_\Delta / \sqrt{10}$, with 9 degrees of freedom. The folds share training records, and selected MusicCaps labels can differ across folds. Therefore, this interval summarizes variation across these ten predetermined folds only. It should not be interpreted as uncertainty from independent retraining or as a confidence-interval coverage guarantee for a broader population. We predeclared a 95% interval and use its lower endpoint as a heuristic cutoff on these fixed folds, not as a population-level significance test. Once the ten fold values are fixed, every decision is deterministic. Our conclusion concerns this candidate rule on these partitions.

3.3 Finite-sample random-label reference

The frequency predictor assigns one constant score to every held-out record. Under our grouped-tie AP calculation, its score is exactly the held-out prevalence k/n. A random strict ranking has a higher exact finite-sample expectation (Bestgen, 2015). An equivalent closed form for k positives among n held-out records and harmonic number $H_n = \sum_{r=1}^{n} 1/r$ is

$$\mathbb{E}[\text{AP}_{\text{strict}}] = \frac{k}{n} + \frac{(n-k)(H_n - 1)}{n(n-1)}. \tag{3}$$

The second term is positive whenever the held-out set contains both classes. A prevalence comparison therefore understates chance-level AP for a strict ranking.

The fitted scorer also produces ties, so we condition the random-label calculation on its observed ordered score blocks. Under a uniform allocation of the k positives across the n records, let block g contain m_g records and let M_g records occur in higher-scoring blocks. Taking the expectation of the grouped-tie AP calculation gives

$$\mathbb{E}[\text{AP}_{\text{tie}}] = \sum_g \frac{m_g(M_g + m_g - 1)(k-1)/(n(n-1)) + m_g/n}{M_g + m_g}. \tag{4}$$

Singleton blocks recover Equation 3; one block recovers prevalence. Appendix A.1 derives the equivalent strict-ranking form and the score-block calculation used here. We compute Equation 4 for every evaluated label and condition, average within each fold, and compare it with the condition score in a separate analysis. This tie-aware comparison was applied after the original audit exposed false acceptances. It is a diagnostic analysis, not a validated candidate decision rule.

3.4 Sensitivity and falsification controls

For MusicCaps, *target masking* removes only the normalized tokens of the candidate label from each caption. A matched-deletion control removes the same number of non-target tokens whenever possible, allowing us to distinguish target-word removal from generic text deletion. Broad masking removes every token appearing in any selected label; because it changes substantially more text, we use it only as a secondary diagnostic, not as a test of the candidate rule.

The two negative controls reassign each record's full multi-label annotation deterministically, separately within each training and held-out split. We call this cyclic complete-label-set reassignment. Source-report fields sometimes call it a label permutation or rotation; it differs from the within-author caption rotation below, which changes captions. For each control, A or B, we form SHA-256 sort keys from a fixed seed and each record identifier. Record i receives the annotation of record $i + 1$ in that ordering, and the last record receives the first record's annotation. We then fit the bag-of-tokens scorer on the reassigned training labels and score the reassigned held-out labels. The controls are two seeded instances of the same operation, not distinct corruption mechanisms. This reassignment breaks record-level text–label pairings while preserving split-level label prevalence. A recipient record and its label donor can have the same vector over the evaluated labels, so not every evaluated vector changes. The integrity audit in Appendix A.10 verifies the donor mapping, change rate, label-set overlap, and prevalence preservation.

MusicCaps also includes a within-author caption rotation, applied separately in training and held-out records. It preserves author-specific language while disrupting record-level caption–label pairing. We apply target masking and the same prevalence reference, but report this separately because it is neither a two-source comparison nor a complete-label-set reassignment.

We also add two sensitivity cues. Exact-label injection adds the target string before masking, so a correct mask should remove its effect. Paraphrase injection assigns each eligible label a phrase with no tokens shared with the label; for example, *acoustic guitar* maps to the synthetic phrase *unplugged sixstring*. At full strength, every positive record–label pair in both the training and held-out sets receives that label's phrase, using one registry across folds. For lower strengths, a fixed random seed selects nested subsets, so every record selected at a lower strength is also selected at higher strengths. The anonymous archive includes the mapping and its digest.

The MTG-Jamendo study follows the same structure with target masking, matched deletion, and two seeded cyclic label-set permutations. A synthetic codeword linked to each target label replaces the paraphrase injection. Within each fold, one deterministic codeword is added to every positive pair for that label, and the identical fold-specific mapping is used in the training and held-out portions. This control tests whether the masked condition can learn and detect label-specific information that does not share a target-label token; it is not evidence of natural-language understanding.

3.5 Data, splits, and reproducibility

Our MusicCaps manifest contains 4,823 public caption and label records from ten authors (Agostinelli et al., 2023). We use leave-one-author-out evaluation and select 16 aspect labels within each training fold. The ten authors define the paired groups. We also report a broad-mask diagnostic with a broader 1,296-label aspect inventory. No waveform enters this analysis.

MTG-Jamendo supplies public genre tags (Bogdanov et al., 2019). Our metadata join retains 55,128 tagged records, assigns artists to ten deterministic hash folds, and selects 16 labels per training fold. Artists do not cross training and held-out portions. This analysis also reads metadata only. For a separate downstream diagnostic, we use the creators' clean music-classification annotations, collected independently from the uploader tags (Music Technology Group, n.d.). They cover three consensus genre taxonomies; 10,598 of the 10,671 annotated tracks join to our metadata manifest. Before reading the human labels, we fixed the three taxonomies and the exact normalized-name matching

procedure. We also fixed the use of the previously fitted scorer, one-versus-rest average precision, the support rule, and an aggregation that gives each taxonomy equal weight across the same ten artist folds. The target-masked scorer was fitted without these human labels. This analysis reuses MTG tracks, so it is not an untouched-source test. It does use separately collected target labels, but it does not measure audio quality.

Post-audit audio-feature protocol. MTG-Jamendo also releases Essentia features extracted from 30-second audio excerpts. We verified the published archive and track checksums and retained 80 mel-band mean and variance coordinates for 55,126 of our tagged records. Before inspecting model outcomes, we fixed a ten-fold artist-held-out comparison. In each fold, we standardized features using the training tracks, fitted one balanced, regularized logistic-regression model per selected genre using either original or reassigned uploader tags, and evaluated the models against separately collected consensus genres on the held-out tracks. Human-consensus labels were never training inputs. We use exact normalized-name matching to link tracks to annotations from four genre taxonomies. A label–taxonomy–fold block enters the primary analysis only if it contains at least five positive and one negative human annotation. We first average labels within each taxonomy, then taxonomies within each fold, and finally folds. We report a paired descriptive Student-t interval over the ten folds. The protocol, source checksums, model settings, fold predictions, and support exclusions are in the anonymous archive. This evaluates how training on genre tags affects prediction; it does not evaluate an existing music-evaluation system.

Post-audit caption-source protocol. We also joined valid human-written Song Describer captions (Manco et al., 2023) to 546 MTG records. Before scoring these captions, we fixed a protocol that selects the five most frequent genres, each with at least 25 training examples. The protocol uses the same ten artist folds, one caption per track, target-label masking, the original text scorer, two complete-label-set rotations, and both prevalence and tie-aware references. Each comparison uses the same fold-level descriptive decision. The source supplies new captions but shares recordings and uploader labels with MTG-Jamendo.

New-collection reference check. Before accessing the Free Music Archive (FMA) metadata (Defferrard et al., 2017), we fixed a check using its balanced 8,000-track small subset, catalog title, artist, and album text, and top-level genre labels. Artists are assigned to ten hash folds. In each training fold, we select the eight most frequent genres with at least 25 examples, then fit the unchanged target-masked scorer. We compare the original pairings and two deterministic cyclic complete-label-set reassignments with both prevalence and the exact tie-aware random-label reference. Decisions use the same ten-fold descriptive lower-endpoint rule. The public source archive passed its published SHA-1 check; the configuration was committed before its contents were inspected. We later found 90 FMA metadata strings that exactly match normalized strings in our earlier MTG manifest. Excluding these records and recomputing the decision from the previously fixed scores is a post-audit sensitivity, not part of the primary check. FMA genre labels and catalog text originate from the same collection, so this test does not measure audio quality or independently annotated genre truth.

Reproducibility. The anonymous supplementary archive `fullsongeval-anonymous-source.zip` contains the scorer, controls, random-reference analysis, genre protocols, audio-feature evaluation, caption-source and FMA checks, reports, fold predictions, source digests, and tests. Its `pyproject.toml` requires Python 3.11 or newer. The README records commands, expected outputs, public-data access steps, checksums, fixed seeds, and implementation revisions for each study. MTG-Jamendo metadata uses CC BY-NC-SA 4.0 terms.

4 Results

4.1 The scorer detects injected metadata cues

On MusicCaps, target-masked captions achieve macro AUPRC 0.1684, compared with 0.0882 for the prevalence reference. The paired difference is 0.0802 with a descriptive 95% summary interval of [0.0500, 0.1104]. This finding shows that caption context retains label-related information after exact target words are removed. It does not show that the information reflects the associated audio.

Appendix Table 10 separates the masking comparisons from the positive-control evidence. Exact-label injection changes target-masked macro AUPRC by only 0.0014 because the injected words are removed. Predefined paraphrases with no label-token overlap increase the same metric on MusicCaps by 0.4675; a distinct synthetic-codeword control increases it by 0.5468 on MTG-Jamendo. Appendix Figure 3(a) shows the contrast across all tested MusicCaps injection strengths. These controls show that the scorer responds to added label-related cues, including cues that do not repeat the exact label text. They do not establish that these cues reflect the audio. The target-masked condition also has lower scores than matched deletion on each source, indicating that target masking removes more label-related information than deleting the same number of non-target tokens.

In a separate MusicCaps broad-mask diagnostic, we compare broad masking with matched deletion under two aspect-label inventories. With the 16 fold-selected labels, the scores are 0.1622 and 0.2150, respectively. With the broader 1,296-label inventory, they are 0.0505 and 0.0580. The same comparison shows no masking effect for two AudioSet-label inventories (Appendix A.9). Thus the information-removal finding depends on the label inventory and is not a general validity test.

4.2 Synthetic-cue sensitivity does not prevent false acceptance

Appendix Table 1 places the scorer check beside the reassigned-label checks using target-masked text throughout. The candidate rule accepts the original MusicCaps captions and labels. It also accepts cyclic label-set permutations A and B. Both controls preserve label prevalence while breaking the record-level pairing between captions and labels. Their acceptance is therefore an error by the rule.

On MTG-Jamendo, the original pairings exceed the prevalence reference by 0.0305 [0.0263, 0.0347]. The two reassignment margins are 0.0019 [0.0012, 0.0026] and 0.0015 [0.0010, 0.0021]; both lower endpoints are positive, so the rule accepts them (Appendix Table 1).

Both MusicCaps reassignment conditions are accepted: their intervals, [0.0046, 0.0228] and [0.0048, 0.0229], have lower endpoints above zero. Appendix Figure 3(b) shows a decreasing but positive margin through complete reassignment. Some selected-label vectors remain unchanged (Appendix Table 11); Appendix Table 12 repeats the comparison only for records whose evaluated vector changed. The two datasets differ in text fields, held-out units, and labels, but share one implementation.

The failure depends on the intervention. For a separate MusicCaps within-author caption rotation, the target-masked margin over the prevalence reference is 0.0360 [-0.0006, 0.0726] (Appendix Table 10), so the rule does not accept that control. The rotation has no matched MTG-Jamendo test.

4.3 The tie-aware reference withholds acceptance for the fixed reassignments

Figure 2 compares the finite-sample references in Equations 3 and 4. The strict random-ranking expectation exceeds constant-score prevalence by 0.0243 on MusicCaps and 0.0014 on MTG-Jamendo; accounting for fitted-score ties changes these values only slightly.

The comparison asks whether the reference represents a plausible null for the scorer's output. For a constant score, grouped-tie AP equals held-out label prevalence, but that score contains no ranking. The fitted text scorer produces a nonconstant ordering even when complete label sets are reassigned. A prevalence reference therefore asks only whether this ordering exceeds a score with no ranking; it can leave a positive margin when the ordering is compatible with randomly placed labels.

Equations 3 and 4 instead evaluate the expected AP under random labels for the relevant finite sample. The strict form treats every score as distinct, whereas the tie-aware form preserves the fitted scorer's score-tie groups. Both references use the same condition scores as the prevalence comparison. Any decision change therefore comes from the null reference, not from refitting the scorer or changing the evaluated records.

This distinction is visible in Figure 2. The observed metadata remains above both references, whereas each fixed reassignment changes from a positive prevalence margin to an interval containing zero under the tie-aware expectation. The result diagnoses a mismatch in the prevalence comparison for these fixed controls; it does not validate the tie-aware reference as a general acceptance rule.

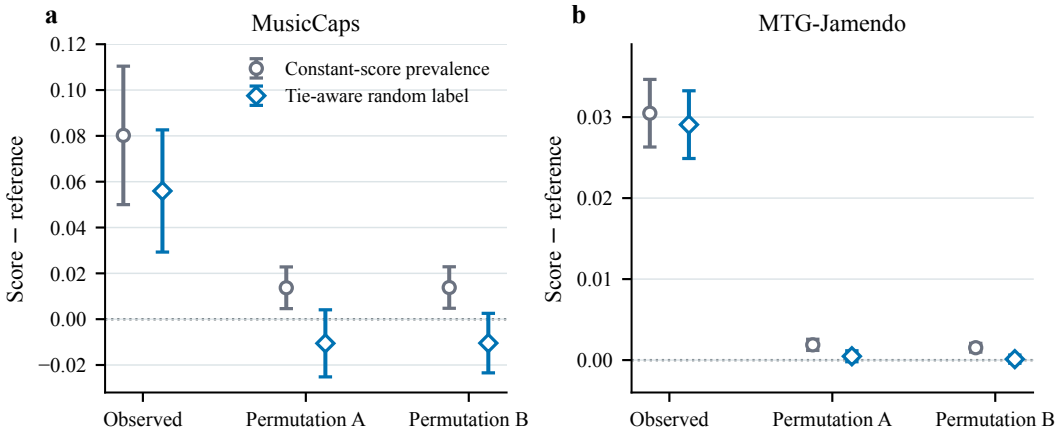

Figure 2: Reference choice changes the fixed-control comparison. Mean score-minus-reference differences and descriptive 95% Student-t summaries are shown for **(a)** MusicCaps and **(b)** MTG-Jamendo. The constant-score prevalence reference places all fixed permutations above zero. The exact tie-aware random-label reference preserves positive separation for observed metadata while every fixed label-reassignment interval crosses zero. The underlying condition scores are identical in both comparisons.

Observed metadata exceeds the tie-aware expectation by 0.0560 [0.0293, 0.0826] on MusicCaps and by 0.0291 [0.0249, 0.0333] on MTG-Jamendo. Every fixed reassignment has an interval containing zero (Appendix Table 2). The score outputs do not change across the two reference comparisons in Figure 2.

In a distinct post-audit check, we independently reallocate the positives for each held-out label, preserving its count while keeping the original scorer outputs fixed. Unlike the two cyclic complete-label-set controls, this check neither preserves multi-label co-occurrence nor refits the scorer. Across 200 draws per source, the prevalence-based rule accepts 124 on MusicCaps and 200 on MTG-Jamendo; the tie-aware reference accepts 0 and 3 (Appendix Table 3). These counts are conditional on the same evaluated groups and should not be interpreted as performance on new groups.

Some reassigned records retain the same selected-label vector (Appendix Table 11). Restricting each control to records with changed vectors still gives positive prevalence margins: 0.0239 and 0.0249 on MusicCaps, 0.0024 and 0.0021 on MTG-Jamendo, with all four lower endpoints above zero (Appendix Table 12). This post-audit restriction changes prevalence; one MTG control also has a small positive tie-aware margin. Thus unchanged vectors alone do not explain the failure, and the tie-aware comparison on the complete held-out folds should still be treated only as a diagnostic analysis.

On 1,300 MARD album reviews (Oramas et al., 2016), both reassigned controls pass prevalence but not the tie-aware reference (Appendix Table 4). The codeword response is too small (AP increase 0.000059) to establish scorer sensitivity, so this is only a reference contrast. Removing cross-fold artist overlap does not change these decisions (Appendix A.4).

4.4 A consequence for an independent genre target

The candidate rule accepts the original MTG scorer and both reassigned controls. We test whether they predict the creators' independently collected consensus genre labels for the same tracks (Music Technology Group, n.d.). The protocol was fixed before the target file was read. Across three taxonomies and ten artist folds, 150 label–taxonomy–fold combinations contain at least one positive and one negative annotation. We compute one-versus-rest AP, averaging labels within taxonomy, then taxonomies within fold, then folds. The consensus labels are never used to fit the scorer.

Original AP is 0.3522, versus 0.2709 and 0.2690 for the two controls. Their paired gaps are 0.0813 [0.0451, 0.1174] and 0.0832 [0.0494, 0.1170] (Appendix Table 5).

Both gaps remain positive in a post-audit sensitivity analysis restricted to blocks with at least five positive annotations. Adding the fourth clean genre taxonomy also preserves them (Appendix A.5). The candidate rule fails to distinguish the original scorer from the reassigned-label scorers on the independent genre labels. The records are previously analyzed MTG tracks, not a test of music quality or an existing evaluation workflow.

In the post-audit audio-feature study, the prevalence-based rule still accepts the target-masked conditions for both original and reassigned uploader tags. On 106 blocks with at least five consensus positives, genre classifiers trained with original tags reach AP 0.7927, versus 0.4365 and 0.4304 with reassigned tags. Paired differences are 0.3562 [0.2972, 0.4152] and 0.3623 [0.2838, 0.4408]; a sensitivity analysis including every supported block yields the same conclusion (Appendix Table 6). Features come from 30-second excerpts on previously studied MTG recordings, not full-song or preference evaluations.

The Song Describer check, whose protocol was fixed before scoring, uses new human-written captions on previously analyzed MTG recordings and labels. The prevalence-based rule accepts observed pairings and both reassignments; the tie-aware reference accepts only the observed pairings (Appendix Table 7).

On 8,000 FMA records, the prevalence-based rule accepts original pairings and both reassignments. The tie-aware reference accepts only original pairings: its AP margin is 0.2471 [0.2081, 0.2861], versus -0.0007 [-0.0029, 0.0015] and -0.0010 [-0.0029, 0.0009] for the controls (Appendix Table 8). A post-audit reanalysis of the fixed scores, after excluding 90 matching metadata strings, preserves all six decisions. FMA is a separate catalog, though its genre labels and text share one source.

5 Discussion

For ranked scores, random-label expected average precision can exceed prevalence, while constant-score average precision equals it. This mismatch can admit reassigned labels without a useful text–label association; it does not explain the independent genre-prediction gaps. The post-audit tie-aware reference accepts original pairings but withholds acceptance for all four primary reassignments and the Song Describer and FMA controls. Song Describer reuses MTG recordings; FMA is a separate catalog. A within-author caption rotation crosses zero even against prevalence, so outcomes depend on the intervention. Prospective audits could report both references, check scorer sensitivity to known cues, and test multiple reassignment mechanisms on untouched data. These results expose a weak rule but do not validate a replacement.

6 Scope and Limitations

We use public text and 30-second excerpts; no experiment measures full-song quality, listener preference, or system comparisons. Synthetic cues do not test natural-language generalization. Song Describer reuses MTG recordings and labels; FMA's overlap check and MARD's artist-overlap sensitivity are detailed in Appendix A.6 and Appendix A.4.

Two cyclic reassignments per source and one bag-of-tokens scorer do not establish behavior for other controls or scorers. The independent per-label allocation holds scores fixed, and caption rotation has no MTG counterpart. The tie-aware reference is conditional on support and ties and was selected after the primary failure. Song Describer and FMA protocols were fixed before scoring but designed after the MTG audit; MARD lacks independent preregistration. We cannot set a general threshold.

7 Conclusion

Above-prevalence average precision does not validate text–tag pairs: the rule accepts four fixed reassignments, whose text and audio scorers predict independent consensus genres less accurately. A post-audit tie-aware reference withholds acceptance for these four, two Song Describer, and two FMA controls but remains diagnostic; future work should test these checks against independent human evaluations.

Reproducibility Statement

Section 3 defines the scorer, controls, and decision rule; Section 3.5 specifies data and folds. The appendix gives derivations and fold results. The anonymous supplementary source archive provides code, configurations, lightweight inputs, checksums, and validation commands.

AI Use Statement

We used generative AI to improve the paper's wording and find related research. We are responsible for the final text and conclusions.

References

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Yves Bestgen. Exact expected average precision of the random baseline for system evaluation. *The Prague Bulletin of Mathematical Linguistics*, (103):131–138, 2015. doi: 10.1515/pralin-2015-0007. URL https://ufal.mff.cuni.cz/pbml/103/art-bestgen.pdf.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The MTG-Jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop at the 36th International Conference on Machine Learning*, 2019. URL https://github.com/MTG/mtg-jamendo-dataset.

Kendrick Boyd, Kevin H. Eng, and C. David Page. Area under the precision–recall curve: Point estimates and confidence intervals. In *Machine Learning and Knowledge Discovery in Databases*, volume 8190 of *Lecture Notes in Computer Science*, pp. 451–466. Springer, 2013. doi: 10.1007/978-3-642-40994-3_29. URL https://doi.org/10.1007/978-3-642-40994-3_29.

Keunwoo Choi, Gyorgy Fazekas, Kyunghyun Cho, and Mark Sandler. The effects of noisy labels on deep convolutional neural networks for music tagging. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):139–149, 2018. doi: 10.1109/TETCI.2017.2771298. URL https://doi.org/10.1109/TETCI.2017.2771298.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. In *Proceedings of the 18th International Society for Music Information Retrieval Conference*, pp. 316–323, 2017. URL https://archives.ismir.net/ismir2017/paper/000075.pdf.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP learning audio concepts from natural language supervision. In *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10095889. URL https://doi.org/10.1109/ICASSP49357.2023.10095889.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. Adapting Frechet audio distance for generative music evaluation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1331–1335, 2024. doi: 10.1109/ICASSP48485.2024.10446663. URL https://doi.org/10.1109/ICASSP48485.2024.10446663.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P. W. Ellis. MuLan: A joint embedding of music audio and natural language. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference*, 2022. doi: 10.5281/zenodo.7316724. URL https://doi.org/10.5281/zenodo.7316724.

Yichen Huang, Zachary Novack, Koichi Saito, Jiatong Shi, Shinji Watanabe, Yuki Mitsufuji, John Thickstun, and Chris Donahue. Aligning text-to-music evaluation with human preferences, 2025. URL https://arxiv.org/abs/2503.16669. ISMIR 2025.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385. Association for Computing Machinery, 2021. doi: 10.1145/3442188.3445901. URL https://doi.org/10.1145/3442188.3445901.

Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6 (4):1–21, 2012. doi: 10.1145/2382577.2382579. URL https://doi.org/10.1145/2382577.2382579.

Ilaria Manco, Benno Weck, Seungheon Doh, Minz Won, Yixiao Zhang, Dmitry Bogdanov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, Elio Quinton, György Fazekas, and Juhan Nam. The song describer dataset: A corpus of audio captions for music-and-language evaluation. In *Machine Learning for Audio Workshop at NeurIPS*, 2023. URL https://zenodo.org/records/10072001.

Music Technology Group. Music classification annotations for the MTG-Jamendo dataset. Official dataset repository, derived music-classification annotations, n.d. URL https://github.com/MTG/mtg-jamendo-dataset/tree/cafd8e20c265ed84f1e61f1c875327971f43a62f/derived/music-classification-annotations. Clean consensus annotation file at source revision cafd8e20c265ed84f1e61f1c875327971f43a62f.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/f2217062e9a397a1dca429e7d70bc6ca-Abstract-round1.html.

Sergio Oramas, Luis Espinosa-Anke, Aonghus Lawlor, Xavier Serra, and Horacio Saggion. Exploring customer reviews for music genre classification and evolutionary studies. In *Proceedings of the 17th International Society for Music Information Retrieval Conference*, pp. 150–156, 2016. URL https://mtg.upf.edu/node/3490.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *The VLDB Journal*, 29(2–3):709–730, 2020. doi: 10.1007/s00778-019-00552-1. URL https://doi.org/10.1007/s00778-019-00552-1.

Takaya Saito and Marc Rehmsmeier. The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432. URL https://doi.org/10.1371/journal.pone.0118432.

Kan Shao. Metadata predictability is not evidence dependence: An intervention-based audit for weak-label benchmarks, 2026. URL https://arxiv.org/abs/2605.23701.

Dapeng Wu, Shun Lei, Wei Tan, Guangzheng Li, Yunzhe Wang, Huaicheng Zhang, Lishi Zuo, and Zhiyong Wu. Songbench: A fine-grained multi-aspect benchmark for song quality assessment, 2026. URL https://arxiv.org/abs/2604.25937.

Li-Chia Yang and Alexander Lerch. On the evaluation of generative models in music. *Neural Computing and Applications*, 32(9):4773–4784, 2020. doi: 10.1007/s00521-018-3849-7. URL https://doi.org/10.1007/s00521-018-3849-7.

Jixun Yao, Guobin Ma, Huixin Xue, Huakang Chen, Chunbo Hao, Yuepeng Jiang, Haohe Liu, Ruibin Yuan, Jin Xu, Wei Xue, Hao Liu, and Lei Xie. Songeval: A benchmark dataset for song aesthetics evaluation, 2025. URL https://arxiv.org/abs/2505.10793.

Jieyu Zhang, Yue Yu, Yinghao Li, Yujing Wang, Yaming Yang, Mao Yang, and Alexander Ratner. WRENCH: A comprehensive benchmark for weak supervision. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/1c9ac0159c94d8d0cbedc973445af2da-Abstract-round2.html.

Table 1: The audit confirms scorer sensitivity, but the candidate rule falsely accepts reassigned labels. All entries use target-masked text. Score gives the arithmetic mean of ten fold-level condition macro AUPRC values; the reference is named in its own column. "Rule accepts" is the candidate-rule outcome for prevalence-reference rows, while "Sensitivity detects" is the separate injection check. Permutations A and B are seeded instances of the same cyclic complete-label-set reassignment. Differences and intervals use unrounded fold values under the full protocol, including fold-specific label selection. The 95% Student-t summary is descriptive because training sets overlap. Blue shading marks the observed-metadata condition; orange shading and bold text mark false acceptances, so the decision does not depend on color alone.

Condition	Reference	Score	Score minus reference (95% t)	Outcome
MusicCaps				
Observed metadata	Prevalence	0.1684	+0.0802 [0.0500, 0.1104]	Rule accepts
Paraphrase addition	Observed	0.6359	+0.4675 [0.4275, 0.5075]	Sensitivity detects
Permutation A	Prevalence	0.1019	+0.0137 [0.0046, 0.0228]	**Rule falsely accepts**
Permutation B	Prevalence	0.1020	+0.0138 [0.0048, 0.0229]	**Rule falsely accepts**
MTG-Jamendo				
Observed metadata	Prevalence	0.0757	+0.0305 [0.0263, 0.0347]	Rule accepts
Codeword addition	Observed	0.6226	+0.5468 [0.5123, 0.5814]	Sensitivity detects
Permutation A	Prevalence	0.0471	+0.0019 [0.0012, 0.0026]	**Rule falsely accepts**
Permutation B	Prevalence	0.0468	+0.0015 [0.0010, 0.0021]	**Rule falsely accepts**

A Expanded Calibration Evidence

This appendix follows the three questions in the main paper: can the text scorer detect a known cue, does the candidate prevalence rule reject mismatched text–label pairs, and what changes when the comparison accounts for finite samples and tied scores? Tables 1 and 10 address the first two questions. The derivation and Table 2 address the third. The remaining sections test the scope of those findings and expose the group-level values behind the summaries. The audio-feature analysis uses statistics from 30-second excerpts; no result measures listener preference or full-song quality.

Several counts below refer to different objects. The primary text studies use 4,823 MusicCaps caption records, held out by caption author, and 55,128 MTG-Jamendo metadata records, held out by artist. Each has ten held-out groups and two seeded cyclic complete-label-set reassignments; the scorer is refitted on the reassigned training labels. The 200 draws per source in Table 3 are a different, conditional experiment: positives are independently allocated per label within each held-out group while the original scores remain fixed. The 546 Song Describer captions share MTG recordings and uploader tags but supply new text. The consensus-genre and audio-feature studies also reuse MTG recordings, with separately collected genre targets; the audio study trains new classifiers on original or reassigned tags. FMA supplies 8,000 records from a separate catalog; MARD supplies 1,300 albums from a separate review corpus. Their ten-fold analyses are distinct from the MusicCaps and MTG folds. None of these counts is a count of independent dataset replications, and a post-audit subset or support restriction does not create a new collection.

Appendix Table 1 gives the complete central calibration comparison used by the main-text discussion. It places scorer sensitivity beside the original and reassigned-label decisions under the same target-masked protocol.

A.1 Derivation of the random-label references

The strict-ranking expectation is known from Bestgen (2015). We give an equivalent closed-form derivation before adapting the calculation to this paper's grouped-tie AP. Let $Y_r \in \{0, 1\}$ indicate whether strict rank r is positive, and let $S_r = \sum_{j \le r} Y_j$. Conditional on exactly k positives allocated

uniformly among n ranks,

$$\mathbb{E}[Y_r S_r] = \frac{k}{n} + (r-1)\frac{k(k-1)}{n(n-1)}. \tag{5}$$

Substituting this identity into $\text{AP} = k^{-1}\sum_{r=1}^{n} Y_r S_r / r$ and using $\sum_{r=1}^{n}(r-1)/r = n - H_n$ gives Equation 3.

For tied scores, let X_g be the number of positives in block g and T_g the positives in all higher-scoring blocks. The grouped-tie implementation contributes $X_g(T_g + X_g)/(k(M_g + m_g))$ for block g. Under the same conditional allocation,

$$\mathbb{E}[X_g T_g] = m_g M_g \frac{k(k-1)}{n(n-1)}, \tag{6}$$

$$\mathbb{E}[X_g^2] = \frac{m_g k}{n} + m_g(m_g - 1)\frac{k(k-1)}{n(n-1)}. \tag{7}$$

Adding these moments, dividing by $k(M_g + m_g)$, and summing over blocks yields Equation 4. We compute this expectation separately for each label, conditioning on that label's observed ordering of tied-score blocks and number of positives. It is not the expected result of the full multi-label cyclic-reassignment procedure.

A.2 Reference-choice comparison

Figure 2 plots the fixed-control score-minus-reference differences, and Table 2 gives the exact values. The constant-score prevalence reference separates all fixed permutations, whereas the exact tie-aware random-label reference does not separate any fixed label-reassignment control. The source scores are identical in both views; only the reference changes.

Table 2: Post-audit comparison with the exact random-label expectation conditioned on each fitted score's tie blocks. Score and reference are mean macro AUPRC over ten held-out groups. The difference and 95% Student-t summary are computed from paired fold values. "Separates" requires a lower interval endpoint above zero. Blue shading marks the observed-metadata condition. Observed metadata separates from the tie-aware reference on both sources; neither fixed label-reassignment control does. This comparison was selected after the original failure was observed, so it is diagnostic, not validation of a replacement decision rule. The rule must be confirmed on untouched data before operational use.

Source	Condition	Score	Tie-aware reference	Difference (95% t summary)	Tie-aware comparison
MusicCaps	Observed metadata	0.1684	0.1124	+0.0560 [0.0293, 0.0826]	Separates
	Permutation A	0.1019	0.1124	−0.0105 [−0.0252, 0.0041]	Does not separate
	Permutation B	0.1020	0.1124	−0.0104 [−0.0234, 0.0025]	Does not separate
MTG-Jamendo	Observed metadata	0.0757	0.0467	+0.0291 [0.0249, 0.0333]	Separates
	Permutation A	0.0471	0.0467	+0.0005 [−0.0002, 0.0012]	Does not separate
	Permutation B	0.0468	0.0467	+0.0001 [−0.0004, 0.0006]	Does not separate

A.3 Independent held-out label allocation with fixed scores

Table 3 asks whether the original acceptance rule also admits chance allocations against a fixed fitted scorer. Within each held-out group and selected label, each draw samples exactly the observed number of positives uniformly over records. We use 200 draws per source, with a separate SHA-256 seed for every held-out group and label. The scorer, score ties, selected labels, and ten evaluated

groups are unchanged. Unlike the primary controls, these draws do not transfer complete label sets or refit the scorer. We compute grouped-tie AP for each draw, average over labels within a group, and apply the same descriptive ten-group interval rule against either prevalence or the exact tie-aware random-label expectation.

The 200-draw mean AP differs from that analytic expectation by 0.000067 on MusicCaps and -0.000005 on MTG-Jamendo. On MusicCaps, the prevalence reference accepts 124 of 200 draws and the tie-aware reference accepts none. On MTG-Jamendo, the corresponding counts are 200 and 3. Thus the prevalence comparison admits many allocations sampled independently of the fixed scores. The Wilson intervals in Table 3 describe uncertainty in these conditional Monte Carlo rates, not new-data uncertainty or the failure frequency of a retrained scorer.

Table 3: Post-audit conditional admission counts from independent per-label reassignment of held-out targets with original scorer scores fixed. Each source has 200 seeded draws over the same ten groups. The 95% Wilson intervals describe Monte Carlo variation in the admission rate conditional on these fixed scores, labels, and groups; they are not population uncertainty. The observed-data comparisons remain positive against both references.

Source	Reference	Accepted/200	Rate (95% Wilson)
MusicCaps	Prevalence	124	62.0% [55.1, 68.4]
MusicCaps	Tie-aware expectation	0	0.0% [0.0, 1.9]
MTG-Jamendo	Prevalence	200	100.0% [98.1, 100.0]
MTG-Jamendo	Tie-aware expectation	3	1.5% [0.5, 4.3]

A.4 Transfer to an album-review corpus

Table 4 reports a separate-corpus check on the Multimodal Album Reviews Dataset (MARD) genre-classification subset (Oramas et al., 2016). Its 1,300 albums each have review text and one of 13 genre labels. We exclude R&B because the unchanged tokenizer produces no maskable token for that label; the other 12 labels are evaluated on ten deterministic album-held-out folds. The local protocol specifies the source file, scorer, masking procedure, fold seed, two cyclic label-set rotations, codeword cue, interval rule, and tie-aware reference. A cryptographic digest of this protocol is recorded in the result report. There is no independent preregistration or timestamped analysis plan created before the run. The rotations preserve every held-out label count and change 93.8% and 91.8% of album labels, respectively. Album identifiers are disjoint between training and test in each fold, but artist identifiers are absent from the classification file.

Observed reviews exceed both references. Each reassigned-label control has a positive lower interval endpoint against prevalence and an interval containing zero against the exact tie-aware random-label reference. The codeword cue increases macro AP over the observed condition by only 0.000059 [0.000042, 0.000076], despite a positive interval endpoint. This small response limits what the sensitivity check establishes for long reviews. The reference contrast transfers to a separately published album-review classification corpus, but the result neither evaluates songs or listeners nor documents an existing workflow using the candidate acceptance rule.

A post-audit match to the creators' full album metadata resolves MusicBrainz artist IDs for all 1,300 albums. The original folds place 41 artists, covering 132 albums, across train and test. In a robustness check, we remove the 40–55 training albums per fold whose artist appears in the held-out fold. The observed tie-aware margin remains 0.0068 [0.0029, 0.0107]. Both rotations still pass the prevalence rule, with margins 0.0283 [0.0239, 0.0327] and 0.0325 [0.0288, 0.0362], while their tie-aware intervals contain zero. This post-audit check addresses artist overlap in training; it does not turn the rule into a validated decision for music quality.

Table 4: MARD transfer on ten album-held-out groups with 12 maskable genre labels. Score is mean target-masked macro AP. Differences use paired descriptive 95% Student-t summaries across groups. The candidate prevalence rule falsely accepts both cyclic complete-label-set reassignments (orange); the tie-aware intervals include zero. The blue condition uses original reviews and genres. Training sets overlap; a separate post-audit check filters artists shared across train and test.

Condition	AP ↑	AP minus prevalence (95% t)	AP minus tie-aware reference (95% t)
Observed	0.1157	+0.0387 [0.0344, 0.0430]	+0.0067 [0.0029, 0.0106]
Rotation A	0.1053	+0.0283 [0.0239, 0.0327]	−0.0036 [−0.0076, 0.0003]
Rotation B	0.1095	+0.0325 [0.0288, 0.0362]	+0.0006 [−0.0032, 0.0044]

A.5 Independent consensus genre target

Table 5 asks whether text scorers accepted by the metadata rule are equally useful for a separately collected genre target on the same recordings. The creators' clean annotation file contains 10,671 tracks; 10,598 match our MTG manifest, but only the tracks with a clean annotation for a given taxonomy enter that taxonomy's comparison. The three selected taxonomies have 621, 573, and 411 annotated tracks, respectively. Exact normalized-name matching yields 150 supported label–taxonomy–fold blocks across the ten artist folds. A block requires at least one positive and one negative human-consensus label, with the same records scored under all three conditions. The original-tag scorer reaches AP 0.3522, versus 0.2709 and 0.2690 after reassignment; the original-minus-control intervals remain positive. Some blocks have only one positive, so a post-audit restriction to at least five positives retains 98 blocks and still yields positive differences of 0.0887 and 0.0913. This comparison shows a consequence of accepted training tags on these MTG tracks, not performance on a new collection; the ten folds share training records.

Table 5: MTG-Jamendo text scorers evaluated against independently collected consensus genre labels on the same tracks. AP ↑ averages matched labels within each taxonomy, taxonomies within an artist fold, then ten folds. Paired differences subtract each reassigned condition from the original scorer. Both reassigned conditions pass the metadata prevalence rule, yet predict the separate genre target less well. The last two rows restrict the post-audit analysis to blocks with at least five positive human annotations.

Condition	Blocks	AP ↑	Observed minus control (95% t summary)
Observed	150	0.3522	—
Permutation A	150	0.2709	+0.0813 [0.0451, 0.1174]
Permutation B	150	0.2690	+0.0832 [0.0494, 0.1170]
Permutation A, $k \geq 5$	98	—	+0.0887 [0.0518, 0.1255]
Permutation B, $k \geq 5$	98	—	+0.0913 [0.0582, 0.1244]

The clean annotation source also contains a fourth genre taxonomy, `genre_electronic`, with 180 annotated tracks; it was omitted from the three-taxonomy primary protocol fixed in advance. A post-audit sensitivity analysis that includes all four taxonomies (supplementary report `reports/mtg_independent_genre_target_all_taxonomies_sensitivity.json`) yields 160 supported blocks and retains positive original-minus-reassignment differences: 0.0777 [0.0473, 0.1081] for permutation A and 0.0799 [0.0441, 0.1157] for permutation B. Requiring at least five positive annotations leaves 106 blocks, with corresponding differences 0.0807 [0.0526, 0.1088] and 0.0784 [0.0523, 0.1045]. Because the analysis reuses the same tracks, it is still not a prospective test on new data.

A.6 Post-audit audio features and new caption source

Table 6 tests whether the genre-prediction gap persists when the human-label support requirement changes. The primary analysis requires five positive human annotations per label–taxonomy–fold

block. Across 106 eligible blocks, the classifier trained on original tags reaches AP 0.7927; classifiers trained on reassignment A and B reach 0.4365 and 0.4304, respectively. Relaxing the support rule to one positive admits 160 blocks. Under this broader rule the corresponding AP values are 0.6834, 0.3876, and 0.3872, and both original-minus-control intervals remain positive. Both comparisons use the same 55,126 feature-complete MTG records and 952 held-out tracks with consensus genres in the selected taxonomies. Published track-level and archive-level checksums verified all 55,699 extracted feature vectors before joining to the tagged manifest. These released audio-derived statistics are not full-song waveforms processed by our evaluator. The support change tests sensitivity to sparse human labels, not a new population or an independent replication.

Table 6: Audio-feature genre task with the primary support rule requiring five positives and a broader sensitivity analysis that includes blocks with at least one positive. AP ↑ is averaged within taxonomy, then artist fold. Paired descriptive 95% Student-t summaries are over ten fixed folds. The first group expands the main-text comparison.

Support	Blocks	Training tags	AP ↑	Original minus control (95% interval)
At least five positives	106	Original	0.7927	—
	106	Reassignment A	0.4365	+0.3562 [0.2972, 0.4152]
	106	Reassignment B	0.4304	+0.3623 [0.2838, 0.4408]
At least one positive	160	Original	0.6834	—
	160	Reassignment A	0.3876	+0.2958 [0.2576, 0.3339]
	160	Reassignment B	0.3872	+0.2962 [0.2471, 0.3453]

Table 7 asks whether new captions on previously studied MTG recordings reproduce the reference contrast. The matched set has 546 Song Describer captions and MTG records. Each selected genre occurs at least 25 times in its training fold; only three to five labels have positive support in each held-out fold. Original pairings have mean AP margins of 0.1152 over prevalence and 0.0494 over the tie-aware reference; both intervals exclude zero. Reassignment A and B also exceed prevalence, with margins 0.1061 and 0.0948, respectively. Neither control separates from the tie-aware reference under the interval rule, despite positive point margins of 0.0404 and 0.0290. The same original or reassigned label counts enter both references within a fold. The broad reassignment intervals reflect the small, uneven held-out groups. This is a new-caption check, not an independent-recording or operational-rule validation.

Table 7: Song Describer caption-source analysis with the protocol fixed before scoring, evaluated across ten artist-held-out folds. Each cell shows mean scorer-minus-reference AP, its descriptive 95% Student-t summary, and acceptance (Y or N). A positive lower endpoint gives Y. Blue shading marks original captions and tags; orange marks controls accepted by prevalence.

Condition	Minus prevalence (95% t); Y/N	Minus tie-aware reference (95% t); Y/N
Original	+0.1152 [0.0872, 0.1432]; Y	+0.0494 [0.0184, 0.0805]; Y
Reassignment A	+0.1061 [0.0217, 0.1905]; Y	+0.0404 [−0.0332, 0.1139]; N
Reassignment B	+0.0948 [0.0153, 0.1742]; Y	+0.0290 [−0.0386, 0.0967]; N

A separate analysis plan, fixed in advance, proposed using Song Describer waveforms but did not meet its support criterion: the matched set yielded only three supported taxonomy–label blocks, so we stopped before fitting an audio model. The audio-feature result above uses the larger MTG release under a separate protocol fixed before model outcomes.

Table 8 gives the new-collection reference decisions and the post-audit overlap check. FMA's small subset contributes 8,000 catalog records with one top-level genre each, distributed evenly across eight genres. The scorer holds out artists across ten folds and masks the tested genre name. Both seeded complete-label-set reassignments pass the prevalence rule with small positive margins but fail the tie-aware reference; the original pairings pass both. Ninety FMA title–artist–album strings exactly match normalized strings in the earlier MTG manifest. Removing those 90 FMA records from the

previously saved score analysis leaves 7,910 records and the same decisions, although the removal can alter per-condition label support after reassignment. This overlap analysis is post-audit and does not change the primary protocol.

Table 8: FMA reference check and overlap sensitivity. Each cell gives the scorer-minus-reference mean AP margin, descriptive 95% Student-t interval over ten artist folds, and the rule's acceptance (Y/N). The first group follows the protocol fixed before the FMA source was inspected; the second removes exact metadata matches to earlier MTG records from its fixed predictions after the primary outcome. Blue shading marks original catalog pairings; orange marks reassignments falsely accepted by prevalence.

Condition	Minus prevalence (95% t); Y/N	Minus tie-aware reference (95% t); Y/N
Primary: 8,000 FMA records		
Original	$+0.2540$ [0.2150, 0.2930]; Y	$+0.2471$ [0.2081, 0.2861]; Y
Reassignment A	$+0.0062$ [0.0044, 0.0081]; Y	-0.0007 [-0.0029, 0.0015]; N
Reassignment B	$+0.0059$ [0.0042, 0.0077]; Y	-0.0010 [-0.0029, 0.0009]; N
After excluding 90 matching metadata strings: 7,910 FMA records		
Original	$+0.2544$ [0.2148, 0.2939]; Y	$+0.2474$ [0.2079, 0.2869]; Y
Reassignment A	$+0.0063$ [0.0043, 0.0082]; Y	-0.0007 [-0.0031, 0.0016]; N
Reassignment B	$+0.0061$ [0.0043, 0.0078]; Y	-0.0009 [-0.0028, 0.0009]; N

A.7 Sources and held-out protocols

Table 9 makes the original two source-specific protocols explicit; the separate MARD protocol appears in Appendix A.4. Label vocabularies are selected from each training fold only. The evaluated-label count can vary for MusicCaps because a selected label must have positive support in the held-out author; all 16 selected MTG-Jamendo labels have held-out support in every fold.

Table 9: Data and original text-protocol summary. No waveform is read in either original text study. The MTG-Jamendo text field concatenates the public track, artist, and album metadata fields. "Min. 25" is the minimum training-fold count required before ranking labels by frequency. The later audio-feature and caption-source checks have separate protocols in Section 3.5.

Source	Records	Text	Target	Held-out unit	Folds	Training-fold selection	Evaluated labels/fold
MusicCaps	4,823	Caption	Aspect labels	Caption author	10	Top 16; min. 25	5–13
MTG-Jamendo	55,128	Track + artist + album	Genre tags	Artist	10	Top 16; min. 25	16

A.8 Scorer-sensitivity, masking, and caption-rotation diagnostics

Table 10 reports the aggregate masking, positive-control, and within-author caption-rotation contrasts cited in the main paper. Figure 3(a) exposes the MusicCaps injection-strength response, while Figure 3(b) shows how the original prevalence-based decision changes as an increasing share of complete held-out label sets is reassigned.

The within-author caption rotation has target-masked macro AUPRC 0.1241 against a 0.0882 prevalence reference, giving a paired difference of 0.0360 with descriptive 95% summary interval [-0.0006, 0.0726]. Although every author-level difference is positive, the interval's lower endpoint is not above zero, so the candidate rule does not accept this control. The complete-label-set false acceptances therefore do not extend to every caption–label disruption; the rotation has no matched second-source test.

Table 10: Masking, positive-control, and within-author caption-rotation diagnostics. Score and reference are mean macro AUPRC across ten held-out groups. The paired difference uses the stated reference. The final column counts held-out groups with positive paired differences. Exact injected label words disappear under target masking, while non-overlapping paraphrases and codewords remain detectable. The MusicCaps caption rotation preserves author-specific language while changing the caption paired with each label set; it is a separate single-source diagnostic, not a complete-label-set reassignment. The masking contrasts address information removal and are separate from the candidate rule's reassigned-label test.

Source	Condition	Score	Reference	Paired difference (95% t summary)	Positive groups
MusicCaps	Target masking	0.1684	0.2230	-0.0547 [-0.0856, -0.0237]	0/10
	Exact-label addition	0.1698	0.1684	$+0.0014$ [-0.0010, 0.0039]	7/10
	Paraphrase addition	0.6359	0.1684	$+0.4675$ [0.4275, 0.5075]	10/10
	Within-author caption rotation	0.1241	0.0882	$+0.0360$ [-0.0006, 0.0726]	10/10
MTG-Jamendo	Target masking	0.0757	0.0768	-0.0011 [-0.0016, -0.0006]	0/10
	Codeword addition	0.6226	0.0757	$+0.5468$ [0.5123, 0.5814]	10/10

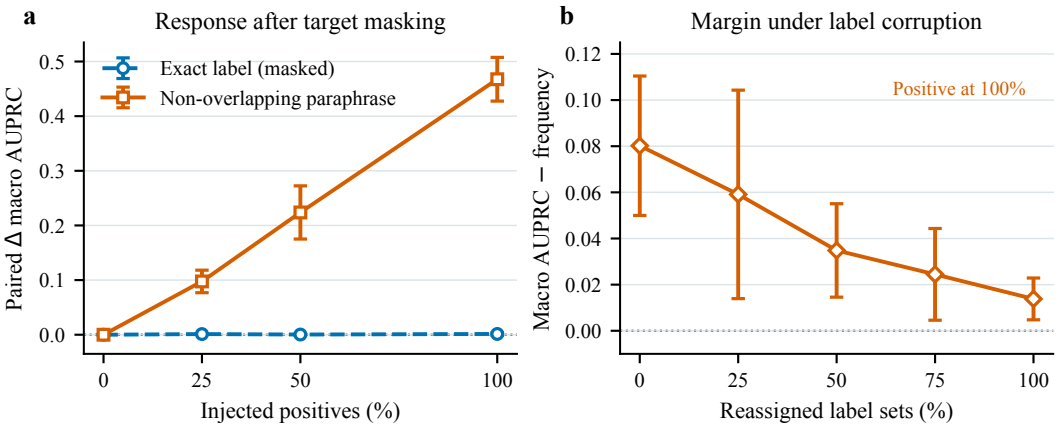

Figure 3: MusicCaps control-response curves over ten held-out caption authors. **(a)** Target-masked macro AUPRC is compared with the no-injection condition. Injected exact label words have negligible effect after masking, while non-overlapping paraphrases produce an increasing response. **(b)** The target-masked score minus the prevalence reference decreases as more complete label sets are reassigned, yet its descriptive 95% Student-t summary remains above zero when all records are selected for reassignment. Some recipients still share their donor's selected-label vector. Error bars summarize the ten fixed author differences. Because the folds share training data, the results should not be interpreted as independent model retraining experiments.

A.9 Completed MusicCaps diagnostics under other label and metric choices

The broad-mask study included four MusicCaps label inventories (supplementary report `reports/musiccaps_stage1_summary.json`). Its deletion control matches the broad-mask token count; the target-matched deletion in Table 10 belongs to the later audit. The main text reports the 16- and 1,296-label aspect inventories, for which broad masking scored below its deletion control by 0.0529 [0.0164, 0.0894] and 0.0076 [0.0032, 0.0120] macro AUPRC, respectively, with the signs here expressed as the control minus broad masking. In the 16- and 211-label AudioSet inventories, normalized unmasked text, target masking, broad masking, and broad-mask-matched deletion gave identical scores within each inventory: 0.2393 and 0.1060, respectively. Thus these AudioSet designs did not show information removal under the implemented masks. The target-mask comparison with broad-mask-matched deletion is descriptive only. None of these four designs tested rejection of reassigned labels.

Other completed MusicCaps text diagnostics use a different prediction task and metric. Their machine-readable summaries are named in the supplementary archive README. In a five-fold top-16 caption-to-tag study, token-model micro-F1 changes from 0.5383 to 0.5064 after label-token masking. A separate hashing-representation test changes from 0.3593 to 0.0728, below its 0.1997 label-prior reference. A distinct top-16 fixed-encoder sweep completed 123 runs spanning 25 encoders and 15 seeds. Its mean raw, masked, shuffled-label, and label-prior micro-F1 scores were 0.1969, 0.1110, 0.1104, and 0.1994, respectively. All 123 masked-minus-prior differences were negative, as were the mean differences for all nine encoder families. The raw mean was already below the prior, so masking alone did not turn an above-reference mean into a below-reference one.

We extended this classification study to eight label-set sizes and 25 encoders, followed by a stability analysis. In the top-1,296-label analysis, all 2,525 masked-encoder results fall below their label-prior micro-F1 reference. An earlier eight-encoder, two-scale sweep at 32 and 64 labels reached the same negative masked-versus-prior conclusion and is subsumed by that broader analysis. These results use different scorers, inventories, splits, and metrics, so they neither validate nor refute the decision made by the candidate macro-AUPRC rule.

The generated-audio pilots and the prompt-matched feature-shift study address a different question. Generator family can be predicted for 189 of 200 raw feature records and all 200 prompt-centered records, where prompt-centered records pair each feature record with the text prompt used for its generation. However, generator identity is not a measure of music quality, listener preference, or the text–label rule studied here. The archive README identifies the pilot feature table and feature-shift report.

A.10 Integrity of the label-reassignment controls

Table 11 verifies what the two negative controls change. Every held-out record receives a donor label set and no record is its own donor. Prevalence is preserved exactly within every fold. The selected-label vector can remain unchanged when recipient and donor share the same selected labels, so the table reports both the change rate and mean selected-label Jaccard overlap. The integrity check therefore distinguishes a successful reassignment operation from the stronger and unnecessary condition that every selected-label vector differ.

Table 11: Integrity audit for the two fixed held-out label-set reassignments, aggregated over the ten disjoint held-out groups per source. "Assignments" counts recipient records. Jaccard overlap is computed after restriction to the labels selected in that fold; two empty selected-label sets receive overlap one. Both controls rotate complete label sets and preserve every selected label's fold-level prevalence.

Source	Control	Assignments	Changed	Jaccard; prevalence
MusicCaps	Permutation A	4,823	3,554 (73.7%)	0.317; preserved
	Permutation B	4,823	3,568 (74.0%)	0.312; preserved
MTG-Jamendo	Permutation A	55,128	47,692 (86.5%)	0.135; preserved
	Permutation B	55,128	47,690 (86.5%)	0.135; preserved

Table 12 repeats the control comparison after selecting only held-out recipients whose vector over the fold's 16 selected labels differs from its donor's. It uses the same scorer fitted on the full reassigned training fold. The subset prevalence and score-tie structure are recomputed, and labels with no positives in that subset are excluded from macro AP. All four prevalence margins remain positive under the descriptive ten-group rule. In this post-audit subset, permutation A on MTG-Jamendo also has a small positive lower endpoint against the tie-aware reference. This subset analysis therefore does not establish that the tie-aware reference is suitable for operational use.

Table 12: Post-audit evaluation restricted to records with changed selected-label vectors. MTG denotes MTG-Jamendo; A and B are the two fixed label-set reassignments. Counts and percentages refer to the complete held-out source; AP differences and descriptive 95% Student-t intervals use the changed subset. Values in the last two columns are multiplied by 10^3. All four prevalence intervals still have positive lower endpoints, so the rule accepts these controls; smaller margins alone do not indicate rejection. The prevalence reference is recalculated on the subset; each scorer remains fitted on its original reassigned training data.

Source	Control	Changed	AP minus prevalence ($\times 10^3$; 95% t)	AP minus tie-aware ($\times 10^3$; 95% t)
MusicCaps	A	3,554 (73.7%)	$+23.88$ [8.31, 39.46]	-8.51 [-21.73, 4.71]
MusicCaps	B	3,568 (74.0%)	$+24.92$ [9.74, 40.10]	-8.87 [-26.77, 9.04]
MTG	A	47,692 (86.5%)	$+2.36$ [1.62, 3.11]	$+0.77$ [0.04, 1.50]
MTG	B	47,690 (86.5%)	$+2.08$ [1.47, 2.68]	$+0.47$ [-0.09, 1.02]

A.11 Corruption strength

Table 13 and Figure 3(b) summarize the same MusicCaps reassignment levels in tabular and graphical form. At each level q, we select records whose deterministic hash value falls below q. Because the same hash values are used at every level, the records selected at a lower q are always included at higher q. Complete label sets are cyclically reassigned within that subset, and the 100% endpoint is permutation B. The mean margin decreases monotonically, confirming that the scorer responds to the reassignment. The operational failure is that the prevalence-based interval remains above zero at every tested level, including complete reassignment.

Table 13: MusicCaps target-masked margin above the prevalence reference as complete label sets are progressively reassigned. Score is macro AUPRC; the paired difference subtracts the unchanged prevalence reference. The decision applies the candidate rule at each fixed reassignment level. Descriptive 95% Student-t summaries use ten held-out authors.

Reassigned label sets	Score	Difference from frequency (95% t summary)	Positive groups	Decision
0%	0.1684	+0.0802 [0.0500, 0.1104]	10/10	Accepts
25%	0.1473	+0.0591 [0.0139, 0.1043]	10/10	Accepts
50%	0.1230	+0.0348 [0.0146, 0.0551]	10/10	Accepts
75%	0.1126	+0.0245 [0.0046, 0.0443]	10/10	Accepts
100%	0.1020	+0.0138 [0.0048, 0.0229]	9/10	Accepts

A.12 Fixed-group contrasts

Tables 14 and 15 expose all held-out-group differences for the observed condition and both fixed label-reassignment controls, and add the positive-cue and masking contrasts. Positive-control differences use the observed target-masked condition as reference; observed and permutation differences use the prevalence reference; masking differences use matched deletion. In Table 14, both permutation margins are negative for Author 6 but positive for the other nine authors, so the false acceptance appears only after the results are aggregated across folds. In Table 15, both permutation margins are positive in every artist fold, though each is at most 0.0039 macro AUPRC. These values describe the fixed protocol and should not be read as ten independent training replications.

Table 14: MusicCaps fold-level paired differences in macro AUPRC. Permutations A and B are the two fixed complete-label-set reassignments. The positive cue is the full-strength non-overlapping paraphrase injection. The final column subtracts matched deletion from target masking.

Held-out group	Observed − frequency	Positive cue − observed	Permutation A − frequency	Permutation B − frequency	Mask − deletion
Author 0	+0.0494	+0.4566	+0.0077	+0.0228	−0.0218
Author 1	+0.0711	+0.5287	+0.0077	+0.0031	−0.0373
Author 2	+0.0621	+0.4485	+0.0292	+0.0191	−0.0620
Author 3	+0.0772	+0.4097	+0.0280	+0.0164	−0.0191
Author 4	+0.1118	+0.5045	+0.0020	+0.0039	−0.1316
Author 5	+0.1649	+0.5374	+0.0286	+0.0427	−0.1328
Author 6	+0.0462	+0.5235	−0.0036	−0.0011	−0.0199
Author 7	+0.1298	+0.3657	+0.0081	+0.0072	−0.0428
Author 8	+0.0309	+0.4396	+0.0259	+0.0136	−0.0271
Author 9	+0.0587	+0.4612	+0.0035	+0.0104	−0.0521

Table 15: MTG-Jamendo fold-level paired differences in macro AUPRC. The positive cue is the fold-specific synthetic codeword injection. Other columns follow Table 14.

Held-out group	Observed − frequency	Positive cue − observed	Permutation A − frequency	Permutation B − frequency	Mask − deletion
Artist fold 00	$+0.0325$	$+0.5130$	$+0.0014$	$+0.0013$	-0.0010
Artist fold 01	$+0.0347$	$+0.5404$	$+0.0010$	$+0.0016$	-0.0026
Artist fold 02	$+0.0223$	$+0.5077$	$+0.0018$	$+0.0019$	-0.0013
Artist fold 03	$+0.0318$	$+0.5375$	$+0.0039$	$+0.0011$	-0.0009
Artist fold 04	$+0.0404$	$+0.5746$	$+0.0026$	$+0.0004$	-0.0016
Artist fold 05	$+0.0247$	$+0.4705$	$+0.0015$	$+0.0014$	-0.0004
Artist fold 06	$+0.0248$	$+0.5909$	$+0.0020$	$+0.0017$	-0.0009
Artist fold 07	$+0.0358$	$+0.5401$	$+0.0016$	$+0.0016$	-0.0009
Artist fold 08	$+0.0260$	$+0.5487$	$+0.0004$	$+0.0011$	-0.0003
Artist fold 09	$+0.0317$	$+0.6448$	$+0.0025$	$+0.0033$	-0.0010