Detector public research

Cross-generator AI music detection

A benchmark report on generalisation, low false-positive operation, and transformation stress

Author Sonny Fullerton
Version 1.0
Source Public repository at commit `4143b39`
Date July 2026
Status Public research report

A compact empirical record for evaluating whether an AI-music detector transfers to generator families excluded from training. The emphasis is not headline accuracy, but performance at a predeclared false-positive operating point and the gap between clean and transformed audio.

https://github.com/sonnyfully/AI-Music-Detection

Executive summary

An AI-music detector is useful only if its decision rule survives contact with new generators, unfamiliar music, and ordinary distribution shift. Random train-test splits do not answer that question: clips from the same generator family can share persistent signatures, while human and generated corpora can differ in codec, loudness, or source pipeline. A model may therefore look excellent while identifying the dataset rather than the origin of the music.

This report records a leave-one-generator-out (LOGO), song-disjoint evaluation of a private reference detector using the public benchmark framework. Each fold removes one generator family from training. A threshold is selected on a disjoint, verified-human calibration set to target a 1% human false-positive rate (FPR), then frozen before evaluation.

0.993
Mean clean ROC AUC

0.87
Mean clean recall at 1% FPR

0.51
Worst clean recall at 1% FPR

0.57
Mean matched-MP3 recall at 1% FPR

The main result is a distinction, not a victory claim. Clean discrimination is strong across ten held-out families, but the matched 40 kbps MP3 condition lowers mean recall from 0.87 to 0.57 at the same operating point. The benchmark therefore supports cross-generator transfer under the tested clean conditions while showing that transformation robustness remains a deployment constraint.

Disclosure boundary. This publication documents the evaluation framework and aggregate reference results. It does not disclose proprietary features, model topology, training recipe, weights, source code, or private data lineage.

1 Question and claims

The benchmark asks: *given a detector trained without a target generator family, how often can it identify outputs from that family while limiting false accusations against held-out human music?*

The evidence supports three bounded claims.

1. The reference system generalises strongly, but unevenly, to ten held-out generator families in the clean evaluation.
2. Performance at a 1% human FPR is materially more informative than aggregate accuracy or F1 for a review-oriented use case.
3. Ordinary matched re-encoding substantially degrades recall and changes which families are hardest.

It does not establish reliable attribution, adversarial robustness, performance on hybrid human-AI works, or calibration under indefinite future drift.

2 Evaluation design

2.1 Unit of generalisation

Generator family is the held-out domain. For each fold, all examples from one family are excluded from training and used only for evaluation. Content-related clips are kept within a single partition so that excerpts or variants of the same song cannot cross the split boundary. Human calibration music is disjoint from both training and test music.

2.2 Operating point

Let $s(x) \in [0, 1]$ be the detector score for audio x. A threshold t is estimated from held-out, verified-human calibration scores. The deployed decision is

$$\widehat{y}(x) = \mathbb{1}\{s(x) \geq t\}.$$

The threshold is chosen once for the target human FPR and then applied without retuning to every held-out generator fold. ROC AUC is reported as a ranking measure; recall at 1% and 5% FPR records performance where a conservative review system would actually operate.

2.3 Controls

The public protocol requires:

- generator-disjoint and song-disjoint evaluation;
- calibration on verified human music excluded from training and testing;
- per-generator sample counts, not only an average;
- both mean and worst-case results;
- matched transformations applied to both human and generated audio;
- explicit disclosure of absent audio, private weights, and incomplete coverage.

These controls are responses to failed early experiments in which source and codec differences created misleadingly high scores. Same-song comparisons and independent human checks became design requirements after those failures.

3 Clean LOGO results

Held-out family	n	Recall @ 1%	Recall @ 5%	ROC AUC
ACE-Step	67	0.10	1.00	1.000
AudioLDM	134	0.98	0.99	0.998
Brev	74	1.00	1.00	1.000
DiffRhythm	139	0.92	0.99	0.994
Mubert	37	0.51	0.97	0.983
Producer	76	0.86	0.97	0.994
SongGen	130	0.99	1.00	1.000
Stable Audio	52	0.54	0.79	0.975
Suno	76	0.97	1.00	0.997
Udio	76	0.89	0.97	0.994
Mean	–	**0.87**	**0.97**	**0.993**
Worst case	–	**0.10**	**0.79**	**0.975**

Table 1: Song-disjoint LOGO evaluation under the clean condition. “Worst case” is the minimum across generator families.

The mean hides an important shape. Eight families exceed 0.50 recall at the strict operating point, while ACE-Step is an extreme low-recall fold at 0.10 despite perfect AUC. That combination is possible because AUC measures ranking across all thresholds, whereas the fixed low-FPR threshold probes only the far human-score tail. It is precisely why AUC alone is inadequate.

Suno and Udio, the two most recognisable commercial families in the set, reach 0.97 and 0.89 recall respectively at the frozen 1% FPR threshold. These are fold-specific measurements, not claims about every model version, export setting, or future release from those providers.

4 Matched transformation stress

Codec and export pipeline can become a spurious shortcut. To test robustness without introducing a human-versus-generated processing mismatch, the benchmark applies the same MP3 transformation to both classes before scoring.

Held-out family	n	Recall @ 1%	Recall @ 5%	ROC AUC
ACE-Step	47	0.68	0.96	0.986
AudioLDM	44	0.73	0.84	0.971
Brev	50	0.66	0.94	0.984
DiffRhythm	49	0.61	0.94	0.984
Mubert	37	0.24	0.70	0.925
Producer	50	0.52	0.94	0.982
SongGen	50	0.96	1.00	0.991
Stable Audio	52	0.37	0.79	0.959
Suno	52	0.63	0.92	0.982
Udio	52	0.33	0.56	0.898
Mean	–	**0.57**	**0.86**	**0.966**
Worst case	–	**0.24**	**0.56**	**0.898**

Table 2: Matched 40 kbps MP3 stress test. Human and generated items receive the same transformation.

The mean 1%-FPR recall falls by 0.30. The family ordering also changes: ACE-Step improves relative to its clean fold, while Udio falls from 0.89 to 0.33. This does not mean compression makes one generator intrinsically

easier or another intrinsically harder; it means the detector's evidence and the calibration tail interact with the transformation in family-specific ways.

5 Human-domain audit

A global threshold can distribute errors unevenly across musical domains. A separate MTG-Jamendo human audit used 1,635 calibration tracks and 865 test tracks. An electronic-music conditional threshold produced 0 observed false positives on that electronic test slice at the 1% target, with approximately 0.5 percentage points of recall cost on the tested Suno/Udio comparison.

This is a measured slice result, not a universal genre guarantee. Conditional calibration spends data quickly: a 1% tail estimate becomes unstable when a group contains only a few hundred calibration items. Every conditional result must therefore report its group sample size and realized test FPR.

6 Interpretation and limits

The clean result is evidence that generator-disjoint evaluation can be passed without disclosing the reference model. The transformed result is evidence that the remaining problem is not captured by a single clean leaderboard number.

Four limits should govern interpretation:

- **Coverage.** Ten families are broad enough to reveal heterogeneity, not to represent all generators.
- **Calibration.** A nominal target is not a permanent guarantee under genre, source, or temporal drift.
- **Stress scope.** Matched MP3 is one ordinary transformation, not an adversarial suite.
- **Decision role.** A detector score is triage evidence. It is not authorship attribution or proof of misconduct.

7 Reproducibility record

The public repository contains a model-agnostic scorer interface, manifest schemas, LOGO orchestration, calibration utilities, leaderboard validation, example submissions, and 29 dependency-light tests. It intentionally contains no audio, detector weights, training code, or private implementation details. Audio manifests retain source and licensing metadata; users must obtain each corpus under its own terms.

The reference result is identified by repository commit `4143b39` and the versioned leaderboard record. A valid independent submission must disclose sample counts, split construction, calibration independence, transformation conditions, and both mean and worst-case metrics.

References

S. Fullerton. *AI-Music-Detection: public benchmark and evaluation framework*. GitHub repository, commit `4143b39`, 2026. https://github.com/sonnyfully/AI-Music-Detection

S. Afchar, G. Meseguer-Brocal, and R. Hennequin. "AI-Generated Music Detection and its Challenges." *arXiv:2501.10111*, 2025.

A. N. Angelopoulos and S. Bates. "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification." *arXiv:2107.07511*, 2021.