Detector public research

Deployment and limitations note

What the Detector benchmark can support, what it cannot, and how to use scores responsibly

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

A risk-oriented interpretation of the public benchmark. This note treats the detector as one signal in a human review workflow, not as an autonomous authorship decision or proof of misconduct.

Decision summary

The public benchmark is strong enough to justify continued evaluation in a controlled review workflow. It is not sufficient to justify autonomous enforcement. Clean cross-generator performance is high, but matched low-bitrate re-encoding lowers mean recall at the strict operating point from 0.87 to 0.57, and human-score tails vary across musical domains.

Review aid
Appropriate role

Not proof
Required interpretation

1% target
Calibration objective, not a promise

Monitor
Thresholds require drift checks

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 Intended use

The detector is designed for triage in licensing, provenance, catalog review, or research settings where a flagged item receives additional examination. A score can prioritise limited reviewer attention or trigger a request for provenance evidence. It should not name a generator provider, identify an author, or independently determine whether a contractual or platform rule was violated.

A responsible interface should present the score alongside the model version, threshold policy, audio condition, and relevant uncertainty. It should use language such as “requires review” rather than “AI-generated.”

2 Why false-positive control is central

In many realistic catalogues, human music is far more common than generated music. Even a low FPR can therefore produce many false flags. If 100,000 human tracks and 1,000 generated tracks are screened at 1% realized human FPR and 80% recall, the queue contains roughly 1,000 false flags and 800 true flags. More than half of the reviewed items would still be human.

This arithmetic is not a forecast; it illustrates why prevalence must be included in product decisions. Precision cannot be inferred from recall and FPR without a deployment base rate.

3 Known limitations

3.1 Uneven human-domain errors

A global threshold can conceal subgroup imbalance. In a public diagnostic slice, measured genre FPRs at the same global policy ranged from low single digits to roughly 10% for some small categories, while the global realized rate was near 1%. Those genre figures are diagnostic rather than stable estimates because the group sample sizes and taxonomy are limited.

Conditional calibration can reduce an observed disparity, but it introduces new failure modes: sparse groups, ambiguous genre assignment, and thresholds that age differently. The correct response is not to treat genre thresholds as permanent; it is to collect larger verified-human calibration banks and monitor each supported domain.

3.2 Transformation sensitivity

The matched 40 kbps MP3 condition reduces mean recall at 1% FPR from 0.87 to 0.57. Re-encoding can weaken model evidence, alter the human calibration tail, and change family ordering. Clean-file scores must not be assumed to transfer to social-media downloads, mastered releases, stems, or transcoded archives.

Every supported ingestion path needs a representative, matched stress evaluation. If the audio pipeline changes, the threshold must be revalidated.

3.3 Calibration resolution

A 1% target lives at the extreme end of the human-score distribution. With a few hundred calibration tracks, the empirical quantile is controlled by only a handful of observations. The threshold may therefore be statistically valid under exchangeability assumptions yet operationally volatile. Larger calibration sets, realized-FPR intervals, and repeated audits are necessary for consequential use.

3.4 Coverage and drift

Ten generator families do not span the space of synthesis methods, provider versions, fine-tunes, or post-processing chains. Some diffusion-heavy families remain difficult in the strict low-FPR regime. Generator updates and changing release practices can move score distributions without notice.

Maintain a temporal challenge set and version every score with the model, calibration bank, and ingestion pipeline used to produce it.

3.5 Unstudied cases

The public result does not establish performance for:

- deliberate adversarial evasion;
- hybrid works combining human performance and generated elements;
- short samples, stems, live recordings, or heavy remixing outside tested conditions;
- attribution to a particular generator or provider;
- legal or contractual conclusions about authorship, consent, or infringement.

4 Recommended review workflow

1. **Validate input.** Record source, duration, codec, sample rate, and any conversion performed during ingestion.
2. **Score and version.** Store the raw score, model version, calibration version, operating threshold, and timestamp.
3. **Apply a triage band.** Route only scores above the declared threshold to review; avoid turning a continuous score into fabricated certainty.
4. **Inspect provenance.** Ask for project files, stems, creation history, export metadata, or other evidence appropriate to the context.
5. **Use a second signal.** Where consequences are material, require independent evidence or a separately developed detector before escalation.
6. **Record outcomes.** Track reviewer disposition, appeals, input domain, and confirmed false positives without silently training on disputed labels.
7. **Recalibrate or stop.** Pause automated routing if realized human FPR, score distribution, or domain mix leaves its validation envelope.

5 Monitoring requirements

At minimum, a deployment log should support the following views:

Monitor	Required question
Input drift	Have codec, duration, loudness, genre, or source distributions changed?
Score drift	Has the human or flagged score distribution moved relative to calibration?
Realized errors	What fraction of reviewed human items were incorrectly routed?
Subgroup performance	Are supported musical domains receiving materially different error rates?
Coverage	Which generators, versions, and transformations are absent from validation?
Appeals	Can a creator contest a flag and supply provenance evidence?
Version lineage	Can every decision be reconstructed from immutable model and threshold metadata?

Realized FPR is usually observable only on a subset with trustworthy labels. Selection bias must be stated: reviewer-confirmed cases are not necessarily representative of all screened audio.

6 Go/no-go gates

Before a production trial, require:

- a licensed and deployment-matched human calibration bank;
- a frozen model and threshold policy;
- matched tests for every supported audio ingestion path;
- enough per-domain human data to estimate realized errors;
- a documented human-review and appeal process;
- audit logs that exclude sensitive audio where retention is unnecessary;
- an explicit stop condition for drift or elevated false positives.

Autonomous rejection, public labeling, payment withholding, or disciplinary action should remain out of scope on the present evidence.

7 Research priorities

The highest-value next work is not another clean average. It is: larger licensed human calibration sets; broader generator and temporal coverage; matched mastering and distribution-chain tests; multi-excerpt track decisions; uncertainty reporting that remains legible to reviewers; and prospective measurement in a low-consequence trial.

8 Evidence and release control

Every public result should be bound to an immutable benchmark record. Updating a model, calibration bank, transformation pipeline, or manifest creates a new version; it must not silently replace the evidence attached to an earlier score. A release note should state what changed, which gates were rerun, and whether the new version is comparable to the prior one.

The present documents describe the benchmark and its aggregate reference result at public repository commit `4143b39`. They are not a model card for a released detector: the implementation, weights, training data lineage, and commercialisation-sensitive details remain private. That boundary allows the evaluation claims to be inspected without presenting the private reference system as reproducible open source.

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.

All benchmark values in this document are reproduced from the public repository at commit `4143b39`.