Framework for Evaluating Audio Quality in Generative Models

Lucas Fernandes Nascimento · May 2026
lucasfnaudio.com | lucasf47@gmail.com

Introduction

Evaluating generative audio is different from evaluating finished records. You are not judging a final artistic product; you are judging a model's attempt to produce one. That means holding two frames simultaneously: is this good audio, and is this a good output given what the model was asked for?

Most evaluation failures happen when these frames collapse into each other. A track can sound impressive while failing the prompt. Another can satisfy the prompt while still carrying technical artifacts, weak arrangement logic, or poor vocal behavior.

This document outlines the working framework I apply across audio evaluation, model-output review, and human-data annotation workflows.

1. Dimensions

I evaluate across five primary dimensions, always in this order.

1.1 Audio Quality — Technical

Clipping, distortion, DC offset, noise floor and SNR, codec signatures such as MP3 swishiness or vocoder artifacts, sample-rate consistency, phase coherence, dynamic range, loudness, and headroom.

1.2 Mix Quality — Engineering

Stereo image and width, frequency balance, masking between elements, dynamic relationships such as sidechain behavior and envelope shape, dry/wet balance, vocal-to-music balance where applicable, and overall mix translation.

1.3 Musicality — Perceptual / Aesthetic

Arrangement logic, melodic coherence, harmonic coherence, rhythmic feel and pocket, emotional arc, catchiness, memorability, and whether the production has a clear musical identity.

1.4 Prompt Adherence — Generative-Specific

Semantic match, style adherence, genre and era accuracy, production-school fidelity, mood or vibe adherence, instrumentation match, and lyrical adherence when applicable.

1.5 Vocal Quality — When Present

Pronunciation and intelligibility, prosody and natural flow, pitch accuracy, vibrato, hallucinated singing or speech, vocal processing choices, breath placement, mouth noise, sibilance, plosives, and whether the vocal delivery fits the intended emotional register.

2. Artifact Taxonomy

Structured vocabulary matters because vague feedback produces weak training signal. Each flag can apply at track level or at a specific timestamp.

Audio Degradation

Clipping, soft clipping, hard digital clipping, harshness, muddy low-mid buildup, vocoder signature, codec artifacts, silence artifacts, end-of-sample static, unstable noise floor, and denoising damage.

Harshness often concentrates around 2–5 kHz. Muddy low-mid buildup often appears around 200–400 Hz. End-of-sample static is a common failure mode in the final seconds of generated outputs.

Mix Problems

Vocal too loud, vocal too quiet, low-end imbalance, missing style-critical elements, dry vocals over a wet instrumental bed, wet vocals over a dry bed, collapsed stereo image, unstable stereo movement, over-compressed master, or elements masking each other.

Musicality Failures

Wrong notes, melody or chord movement outside the intended key, awkward arrangement, weak section lengths, abrupt transitions without musical purpose, flat dynamic arc, generic identity, and emotional mismatch.

Prompt Failures

Genre mismatch, era mismatch, instrumentation mismatch, production-style mismatch, lyrical drift, wrong emotional register, missing requested element, or adding an element that contradicts the prompt.

Vocal Failures

Bad pronunciation, missing lyrics, hallucinated lyrics, unnatural phrasing, pitch failures, register flips without musical reason, robotic breath placement, or vocal processing that fights the intended style.

Intro Failures

Weak commitment in the first 3–5 seconds, intro inconsistent with the rest of the track, immediate quality issue on entry, or an opening texture that contradicts the requested style.

Excellent

A reserved positive tag for outputs that stand out in quality, catchiness, prompt adherence, or all three. I use this sparingly; if everything is excellent, the tag stops meaning anything.

3. The Hardest Judgment: Intent vs. Failure

The core discipline of this work is distinguishing stylistic choices from technical failures. A bad take by a great artist is still a bad take. A "mistake" that defines a genre is not a mistake.

I apply four principles.

3.1 Genre Baseline First

Before flagging anything, establish what the genre accepts as normal.

Lo-fi hip-hop accepts low SNR as aesthetic. Death metal accepts harshness as texture. Trap accepts extreme autotune as central to the form. Dub and ambient can accept long tails, hiss, and blurred transients. If it is expected in the genre, it is not automatically a failure.

3.2 Consistency Test

If an artifact appears throughout with intention, it is probably stylistic. If it appears once, unexpectedly, and breaks continuity, it is probably a failure.

A filtered vocal across an entire intro is a production choice. One random muffled phrase in an otherwise clear vocal is more likely an artifact, edit problem, or model failure.

3.3 Engineer Test

Would a professional engineer release this take?

If yes, it is likely intentional or at least acceptable within the style. If no engineer would let it pass, it is likely a failure. This test is especially useful for clipping, breath noise, vocal level, unstable low end, and harshness.

3.4 Common Ambiguous Cases

- **Autotune** — stylistic unless it visibly fights the source.
- **Distortion** — stylistic unless it clips the master bus or destroys intelligibility.
- **Lo-fi character** — stylistic unless SNR prevents comprehension.
- **Pitch drift** — stylistic in slowed-and-reverb contexts, failure in clean pop.
- **Dry vocals** — stylistic in punk or indie, often a failure in polished R&B.
- **Compressed dynamics** — expected in electronic music, often wrong in jazz or classical.

When in doubt, I flag with context: "harshness around 3 kHz — likely stylistic given genre baseline, but worth checking against reference."

4. Timestamp Description Methodology

For section-level annotation, I follow a consistent structure per segment:

1. **Lead element** — what dominates the section.
2. **How it is played** — behavior of the lead element.
3. **Groove** — backbeat, feel, pocket, rhythmic engine.
4. **Supporting elements** — bass, pads, effects, background vocals, percussion.
5. **Texture** — minimal, layered, building, reduced, dense, sparse.

6. **Mood** — emotional register.
7. **Change marker** — what shifted from the previous segment.

Example Applied
00:45–01:02 — Transition

- Male vocal, mid pitch, speech-like delivery, drops out mid-segment.
- Groove reduces to a kick-only pulse.
- Sub-bass sustains a long note.
- White-noise sweep rises across the section.
- Texture is sparse and anticipatory.
- Mood is tense and suspended.
- Change: vocals and harmonic content drop; sweep signals the incoming drop.

This format produces training-ready data: structured, consistent, and specific enough that another annotator can verify the labels without needing my interpretation first.

5. Preference Scale for A/B Evaluation

For comparative evaluation of two outputs from the same prompt, I use a scale that separates strength of preference from the reason behind it.

- **Strongly Prefer A** — A is clearly better across multiple dimensions.
- **Prefer A** — A is better overall, but B has redeeming qualities.
- **Equal** — comparable overall quality; notes specify dimensional trade-offs.
- **Prefer B** — inverse of Prefer A.
- **Strongly Prefer B** — inverse of Strongly Prefer A.

When marking **Equal** and both outputs are weak, I specify: *Equal — both below threshold*. When both are strong, I specify: *Equal — both above threshold*.

Strongly Prefer should remain rare. It is reserved for cases where one output is clearly correct and the other is clearly wrong. Overuse collapses the scale and reduces the usefulness of the evaluation signal.

6. Feedback Formatting

Structured feedback is more useful than freeform critique. Each evaluation should produce:

- Tags applied from the artifact taxonomy.
- Timestamp notes for section-specific issues.
- Overall verdict in one clear sentence.
- Intent-vs-failure annotations where the case is ambiguous.
- Suggested reference when a known recording illustrates the intended direction.

The goal is feedback that another annotator can replicate, that an ML engineer can act on, and that produces consistent training signal across reviewer pools.

Closing

This framework is not exhaustive. It is the working discipline I apply across audio evaluation, model-output review, and human-data annotation workflows — tightened across a year of evaluating generative outputs and annotating human-made reference music for training pipelines.

It evolves as models evolve. What I flag as a failure in a 2025 model may become baseline by 2027.

The discipline underneath it is older than the AI use case: fourteen years of producing, mixing, performing, and criticizing music. The framework simply gives that ear a structure.