

AI Music Generation: Evaluation Guidelines

A rubric, failure taxonomy and listening protocol for scoring AI-generated music clips consistently, so two raters (or the same rater five or more days later) reach the same score for the same reasons.

Author: Adam Asifo **Applies to:** 20–60 second text-to-music clips **Companion file:** AdamJAsifo_AI_Audio_Scoring_Workbook.xlsx

1. What a rater produces for every clip

1. A 1–5 score on each of six dimensions, using the written anchors in section 3.
2. Up to three failure codes from the taxonomy (section 5), with the timestamp of the first occurrence.
3. A hard-fail flag where section 4 applies.
4. One or two sentences of rationale saying what was heard, where, and why it cost points (section 7).

2. Listening protocol

Setup (write this down once per session)

- Closed-back headphones at a fixed, moderate volume. Record the model.
- Second check on laptop speakers for the mix and artifact dimensions.
- Loudness-match clips before comparing. Louder always sounds better, and that bias is the most common scoring error.
- Quiet room; no other audio playing.

Per clip

- **Pass 1:** full listen, no notes. Judge prompt adherence, structure and groove.
- **Pass 2:** listen for defects. Pause and timestamp each one.
- Score only after both passes. Don't re-score after reading other raters' notes.
- Take a 5-minute break every 10 clips; ear fatigue inflates artifact scores.

3. Rubric and anchors

Dimension	Weight	1 — Unusable	3 — Acceptable with flaws	5 — Release-ready
Prompt adherence	20%	Ignores the prompt: wrong genre, tempo or mood.	Right genre and mood; one requested element missing or off.	Every requested element present: genre, tempo, mood, instruments, vocal type.
Structure & arrangement	15%	No recognizable sections; ideas start and stop at random.	Clear sections, but one transition is abrupt or a section overstays.	Sections build and resolve; transitions feel intentional.
Mix balance & clarity	15%	Key elements buried or overpowering; masking everywhere.	Balanced on headphones; one element masks another in busy passages.	Every element has space; translates on headphones and laptop speakers.
Artifacts & fidelity	25%	Artifacts obvious even on laptop speakers.	Artifacts audible only on close headphone listening, in 1–2 spots.	No audible artifacts on close headphone listening.
Vocal intelligibility	15%	Lyrics unintelligible; syllables collapse.	Most words clear; consonant clusters or fast passages blur.	Every word intelligible at normal volume; natural phrasing.

Timing & groove	10%	Tempo drifts or parts fall out of time.	Mostly steady; one section rushes, drags or loses the pocket.	Locked groove from start to finish.
----------------------------	-----	---	---	-------------------------------------

Scores of 2 and 4 sit between anchors. On instrumental prompts, **vocal intelligibility is left blank** and the other weights rebalance automatically. **Weighted score** = $\Sigma(\text{score} \times \text{weight}) \div \Sigma(\text{weights used})$. A clip **passes** at a weighted score of 3.5 or higher with no hard fail.

4. Hard-fail rules (the clip scores 0 regardless of other dimensions)

- Silence, dropouts, or the audio cuts out for more than one second mid-clip.
- Clipping or distortion across the whole clip (not just genre-appropriate saturation).
- Completely wrong genre, e.g. an orchestral piece for a drill prompt.
- The clip closely imitates a specific real song's melody or a specific real artist's voice. Flag it for the project lead. The note describes the resemblance and doesn't guess at identities.

5. Failure taxonomy

Code	What it sounds like	Don't confuse it with
F01	Transient smearing Kicks and snares lose their attack; hits sound soft or 'washed'.	A deliberately soft, low-passed lo-fi drum sound the prompt asked for.
F02	Metallic / phasey high end Hi-hats and cymbals swirl, chirp or sound 'underwater'.	Real phaser or flanger effects that fit the genre.
F03	Vocal phoneme collapse Words blur into vowel mush, especially clusters like 'str', 'nts', 'sk'.	Stylized mumble delivery that stays rhythmically consistent.
F04	Tempo or timing drift The groove speeds up or slows down, or parts fall out of time with each other.	Intentional swing or a written tempo change.
F05	Structural incoherence Sections don't connect: a hook appears once and vanishes, or a verse cuts off mid-bar.	A hard cut at the generator's time limit (see edge-case rule 4).
F06	Muddy low end / masking Bass and kick fight; low-mids cloud the vocal.	Heavy 808s that are loud but still defined.
F07	Unintended clipping / distortion Crackle or crunch on peaks across multiple elements.	Distorted 808s or guitars that are normal for the genre (rule 2).
F08	Stereo image problems Mix collapses to mono, leans hard to one side, or elements jump sides.	Intentional panning moves.
F09	Prompt miss: genre or mood Asked for gospel and got pop; asked for dark and got bright.	A reasonable sub-genre interpretation.
F10	Prompt miss: instruments or vocals Asked for instrumental and got sung lyrics; a named instrument is missing.	Wordless vocal chops used as texture (rule 3).
F11	Artifact tail Phantom reverb, noise or a stray syllable at the very start or end.	A natural reverb tail on the last chord.

F12	Repetition loop The clip gets stuck repeating one phrase or bar.	A repeated hook that develops (new words, added layers).
-----	--	--

6. Edge-case rules

- Requested lo-fi is not an artifact.** If the prompt asks for vinyl crackle, tape hiss or a muffled sound, score fidelity only on problems the prompt didn't ask for.
- Genre-normal distortion.** Saturated 808s in trap and drill are expected. Tag F07 only if the distortion spreads to other elements or crackles at peaks beyond what the genre expects.
- Vocal chops in an instrumental.** Wordless chops used as texture are fine. Sung or rapped lyrics on an instrumental prompt are F10, and prompt adherence can't score above 3.
- Generator time limits.** Don't penalize structure because the clip hard-stops at its length limit. Judge the structure of what's inside the window.
- Gibberish or wrong-language lyrics.** Vocal intelligibility = 1 and tag F03, even if it sounds pleasant.
- Torn between two scores.** If the flaw is audible on laptop speakers, take the lower score. If it only shows up on close headphone listening, take the higher one.
- Several failures at one moment.** Tag all of them (up to three) and timestamp the first. If more than three apply, keep the three that cost the most points.
- Can't tell whether it's the model or your playback chain.** Re-listen on the second device. If it's still present, it's the model's; if not, note it and don't tag it.

7. Writing the rationale

A rationale is useful to a model team only if it says **what** happened, **where**, and **which dimension** it hit. The examples below are illustrative.

Weak	Strong
"Drums sound bad and the vocals are kind of hard to understand."	"Snare loses its attack from 0:12 on (F01, fidelity 3). Chorus lyrics blur on 'streets' and 'lights' at 0:21 (F03, vocals 2). Groove holds steady."
"Doesn't really match the prompt."	"Prompt asked for an instrumental; a sung hook enters at 0:18 (F10). Genre and tempo are correct, so prompt adherence gets a 3, not a 1."
"Good song, 5s across the board."	"No artifacts on headphones or laptop. Hi-hats are clean and the low end is defined. One small mix issue: the pad masks the vocal in the bridge (0:34), so mix gets a 4."

8. Staying consistent (and proving it)

Protocol: Five or more days after the first session, re-score 10 clips without looking at the original scores. The workbook's Consistency_Check tab compares the two sessions.

Measure	What it tells you	Target
Exact agreement	Share of dimension scores that came out identical	≥ 60%

Within ± 1	Share that differ by no more than one point	$\geq 95\%$
Quadratic weighted kappa	Agreement beyond chance; big disagreements count more than small ones	≥ 0.70
Drift by dimension	Average change per dimension. A consistent + or - means the anchor is being read differently.	within ± 0.25

If a dimension drifts beyond ± 0.25 , re-read its anchors, write a one-line calibration note in the log, and re-score the affected clips.

Appendix: lyric transcription conventions

Built from preparing lyrics for a 15-song distribution project, where the hardest part was keeping dialect spellings consistent across every song.

- **Transcribe what's performed.** Don't correct grammar or standardize dialect. If the artist says *finna*, the lyric says *finna*.
- **One spelling per word, per project.** Keep a spelling list (e.g. *trynna*, *finna*, *gon*, *shi*) and apply it the same way in every song. Mixed spellings of the same word are the most common error.
- **Check the list before delivery.** Search each file for variants of every word on the list before submitting.
- **Mark uncertainty; don't guess.** If a word is unclear after three listens, flag it with its timestamp for the artist to confirm.

Written by Adam Asifo. Claude (Anthropic) assisted with drafting and formatting. The rubric, anchors and edge-case rules reflect my own listening approach from audio engineering, studio work and music production.