MEASURING CONTEXT USE IN MUSIC GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

How can we measure whether a musical continuation carries forward one player's part and relates to the other players? We introduce Predictive Musical Information (PMus), a framework grounded in information theory for evaluating a continuation's relationship to its shared context and individual players. PMus uses an autoregressive model as an evaluator to compare a fixed continuation's predictability when parts of its past context are provided, withheld, or replaced. It accommodates the single mixed output common to music generators, without requiring separate tracks for each player. With an external evaluator, PMus can assess human performances and outputs from open and closed source generators without access to the generating model's parameters or probabilities. We evaluate five measures across four open music generation models through controlled experiments and a listening study with 85 musically trained participants. The context contrasts agree with musicians' choices on 62–67% of judgments, with the shared-context measure PMus-Duo highest at 66–67%, whereas ranking by predictability under the true past alone agrees below chance on the same judgments. The measures also respond differently to controlled changes in what the generator was given. Because listeners answered the two questions alike, the perceptual separation of self-continuation from partner response remains open. PMus provides a practical foundation for evaluating how generated music relates to its context.

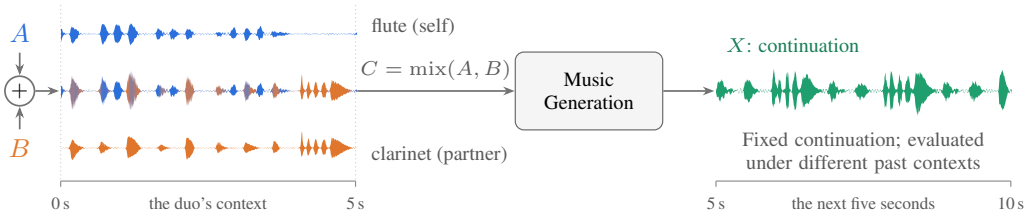


Figure 1: **What does the musical past contribute to a continuation?** Players A (flute) and B (clarinet) share a five-second past, mixed as $C = \mathrm{mix}(A, B)$. A model generates X, intended to continue A. PMus then holds X fixed and measures how its predictability changes when parts of the past are withheld or replaced. These comparisons assess predictive contributions from the shared context and each player's past. Mixture colors indicate each player's relative short-time energy.

1 INTRODUCTION

A musical passage can sound convincing on its own yet fit poorly with what came before it. Evaluating music generation therefore requires more than judging the output in isolation. We must also examine its relationship to the musical context. When that context contains several players, the question becomes more specific: how does the output relate to each player's contribution?

Existing methods assess accompaniment coherence, stem compatibility, and adherence to an audio prompt (Ciranni et al., 2025; Riou et al., 2024; Grachten & Nistal, 2025). We ask a complementary question: how does a fixed continuation's predictability change when part of its past context is removed or replaced? For a duo, this lets us compare what the shared past contributes and what each player's past adds beyond the other's. Throughout, the *target player* A is the one whose part is

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.

continued and the *partner* B is the other player; the self measures concern A's past and the partner measures B's.

We introduce Predictive Musical Information (PMus), a protocol for making these comparisons. PMus builds on information dynamics in music and measures of contextual specificity in natural language processing (Gokul et al., 2024; Paranjape & Manning, 2021). In music with several simultaneous parts, evaluating the combined context alone does not tell us what each player's past adds to the predictability of the continuation. Quantifying these reciprocal influences matters for interactive agents, which must continue their own part while incorporating the other sources; a per-player measure can serve as a training or selection signal for them. PMus uses a probabilistic model to evaluate an already generated continuation. For example, *PMus-Self*, one of our five measures, compares the probability this model assigns to the same sequence of continuation audio tokens when given both players' pasts and when given only the partner's past. The difference in average log probability measures how much adding the target player's past helps the model predict that fixed continuation. The scoring model can be separate from the music generator, so the generator's own probabilities are not required.

We study five measures that address three questions: does the shared past help predict the continuation, does the target player's past help beyond the partner's, and does the partner's past help beyond the target's? Across four music generation models, controlled experiments test sensitivity to generation context, while a listening study tests agreement with musicians' judgments of contextual fit.

Our contributions are:

- **Predictive contrasts for audio continuations.** PMus carries pointwise context contrasts from dialogue and symbolic music to audio-token models, measuring what the shared past and each player's past add to a continuation's predictability. The measures accommodate a single mixed output without requiring its separation into player-specific tracks.
- **Evaluation across generators.** PMus uses an external probabilistic model to assess contextual fit without access to the generating model's parameters or probabilities. This allows outputs from open and closed generators, as well as human performances, to be evaluated under a common scoring protocol. We test this approach across four generators and examine how the choice of evaluator affects listener agreement.
- **A validation protocol.** Controlled changes to what a generator is given, paired with a listening study of 85 musicians, test the measures' intended readings. To our knowledge this is the first listener validation of an information-dynamics measure of musical interaction at this scale.
- **Findings.** The context contrasts agree with musicians where raw predictability does not; partner access moves the partner measures and not the self measure; and listener alignment depends on the evaluator. Perceptual separation of self and partner is not established.

2 Related Work

From dialogue to musical interaction. Pointwise mutual information (PMI) is established in NLP, from lexical association to dialogue generation (Church & Hanks, 1990; Li et al., 2016). With several context sources, overall relevance can hide dependence on only one. Paranjape & Manning (2021) use pointwise conditional mutual information (PCMI) to assess what conversational history adds beyond supplied facts. Nandwani et al. (2023) apply a related comparison to document-grounded faithfulness. These applications motivate our approach, but its musical interpretation requires testing. In music, players communicate simultaneously and share rhythm, harmony, and acoustic space. Their contributions overlap, and the full audio context mixes them rather than identifying each source separately.

Information dynamics in music. Dubnov (2006) studies predictive structure in MIDI and audio through information rate. Later work examines musical interaction through transfer entropy (Nakayama et al., 2020) and Total Information Flow (Gokul et al., 2024). The latter uses a symbolic multitrack Transformer to evaluate musical co-creativity. Two gaps separate this work from ours: the measures operate on symbolic streams rather than audio tokens and they use the generating model's own probabilities. PMus addresses each: audio-token contrasts with fixed target encoding, an external evaluator, and a listening study. Whereas Total Information Flow treats players as separate

musical streams, PMus accommodates generators that return a single mixed audio output, without requiring its separation into player-specific tracks. Audio carries additional detail, including timbre and articulation (Dubnov, 2021). Gains in predictability can therefore reflect acoustic continuity as well as musical structure.

Musical expectation. Our interest in predictability as a possible indicator of musical fit is also motivated by research in music cognition on how listeners form expectations from learned musical regularities (Meyer, 1956; Huron, 2006). Pearce's IDyOM framework makes this connection computational: it learns musical regularities and expresses surprisal as negative log conditional probability (Pearce & Wiggins, 2012). Behavioral and neural evidence links its predictions to musical expectation (Pearce et al., 2010). These findings motivate our test of whether contextual changes in audio-token surprisal track musicians' judgments.

Audio-context evaluation. Existing methods assess how well musical parts fit together. A stem is a separate audio track containing one musical part or a group of parts. COCOLA scores coherence between stems (Ciranni et al., 2025), while Stem-JEPA predicts compatible stem embeddings from a context mix (Riou et al., 2024). Accompaniment Prompt Adherence evaluates distributions of context–stem pairs against matched and mismatched references (Grachten & Nistal, 2025). Our focus is a future continuation and its relationship to each player's past. We score the same clip with the shared past, with individual parts withheld, and with unrelated parts substituted. This separates overall context gain from what each player's past adds beyond the other's in the evaluator's predictions. We test these comparisons through generation interventions and listener judgments. The former reveal different responses across measures; the latter support contextual fit.

3 Predictive Musical Information

3.1 Mathematical Foundation

Let A and B denote the target player's and partner's past audio, with shared context $C = \mathrm{mix}(A, B)$. For fixed continuation tokens $X = (x_1, \ldots, x_n)$ and a conditional probability function p_θ, used as the evaluator, define

$$h(X \mid C) = -\frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x_i \mid x_{i-1}, \ldots, x_1, C) \tag{1}$$

We hold the evaluator parameters θ fixed within each context comparison and suppress this dependence in h for conciseness; n is the length of the fixed token sequence X. Negative log probability measures surprise. Each term describes how unexpected a token is given the context and preceding continuation tokens. Their sum measures whole-continuation surprisal; dividing by n gives an average in nats per token. Lower values mean greater predictability. Scoring uses actual preceding tokens, not sampled predictions (*teacher forcing*).

PMus compares this score for the same continuation under different contexts. A decrease in mean surprisal indicates that the supplied context makes the continuation more predictable. Section 3.2 defines these comparisons.

3.2 Context Measures

Table 1 groups five measures by shared and player-specific context. References withhold a part or replace it with an unrelated recording (a decoy). *PMus-Self* and *PMus-Partner* measure what each player's past adds beyond the other's, not separate shares of the total context gain.

Table 1: **PMus context measures.** Each score measures the reduction in mean surprisal of a fixed continuation X under the true shared past C, relative to a reference context. Scores are expressed in nats per token; positive values indicate greater predictability under C. See Figure 2 for an example.

Measure	Definition	Interpretation
Shared context		
PMus-Ctx	$h(X \mid \emptyset) - h(X \mid C)$	Predictive gain from the true shared past relative to no audio context.
PMus-Duo	$h(X \mid \bar{C}) - h(X \mid C)$	Predictive gain from the true shared past relative to unrelated duo pasts.
Target-player context		
PMus-Self	$h(X \mid B) - h(X \mid C)$	Predictive gain from the shared past relative to the partner's past alone.
Partner context		
PMus-Partner	$h(X \mid A) - h(X \mid C)$	Predictive gain from the shared past relative to the target player's past alone.
PMus-Partner$^+$	$h(X \mid \mathrm{mix}(A, \bar{B})) - h(X \mid C)$	Predictive gain from the true partner relative to unrelated partners, with the target player's past retained.

$C = \mathrm{mix}(A, B)$; $\emptyset$ denotes no audio context, and bars denote unrelated substitutes. The primary MusicGen evaluator averages decoy contrasts over eight substitutes.

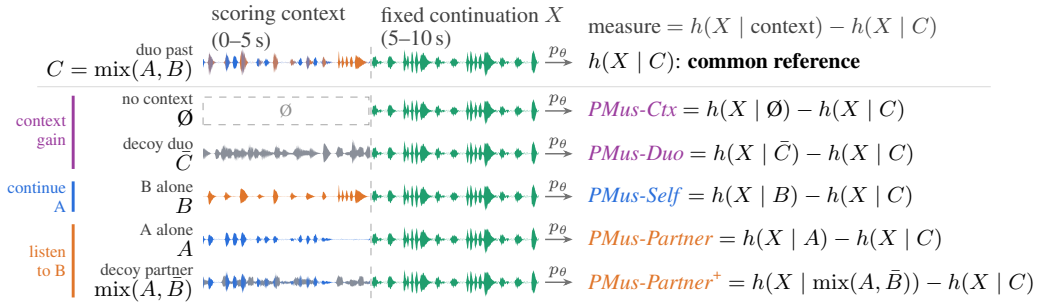


Figure 2: **Computing PMus from a fixed continuation.** The evaluator changes scoring contexts, not X; no new continuation is generated. Rows illustrate Table 1 using the example in Figure 1. One of eight scoring decoys is shown. Flute blue, clarinet orange, decoy grey; mixture colours show each source's share of energy.

Withholding or replacing context. Withholding requires no unrelated recordings. Decoys test context specificity but require a reported decoy set and additional scoring passes. Neither approach reliably outperformed the other with listeners (Appendix D.4); their reference contexts and computational costs favor different use cases.

Interpretation. The intuition behind PMus is that a musical continuation's relationship to its past is reflected in how that past helps predict it. Predictive gain quantifies this relationship under an evaluator: shared-context measures capture what the ensemble's past contributes, while player-specific measures capture what one player's past adds beyond the other's. The controlled experiments and listening study test how these gains relate to generation context and perceived musical fit.

Two cases merit care when reading the scores: noise can be surprising under both contexts yet yield little gain, and players carrying similar musical information can yield small incremental gains even when their musical relationship is strong (Appendix A).

4 EXPERIMENTAL SETUP

4.1 MUSICAL DUO DATASET AND GENERATORS

Figure 1 illustrates the setting. The corpus has 44 duos with separate stems: 11 from URMP (Li et al., 2019), 21 from Slakh (Manilow et al., 2019), 8 from MedleyDB (Bittner et al., 2014) and 4 separated from MusicGen (Copet et al., 2023) output. Curation aimed to select two individual musical lines rather than a lead melody over accompaniment. Collection-specific procedures are detailed in Appendix B. Each track gives one to three five-second context windows in which both players are active (116 windows). An item is a window with one of its players as the target A and the other as the partner B; both directions are kept, giving 232 items, each with five seconds of past and five seconds of source future. Four generators continue A from the past: MusicGen-large (the past mix as decoder prefix, a text prompt naming A's instrument), InspireMusic-1.5B (Zhang et al., 2025) (the past mix as audio prompt), Magenta RT (Caillon et al., 2025) (A's past as token context, B through its style embedding) and ACE-Step (Gong et al., 2026) (turbo, guided halfway between A alone and the mix). YuE (Yuan et al., 2026) is included only in the pilot analyses reported here. Appendix B also gives window selection, conditioning, sampling settings, seeds and coverage.

4.2 ONE-INSTRUMENT SCREENING

The continuation task asks for one player, but generators often add additional players nonetheless. In the pilot experiment, human listeners reported 115 of 350 continuations as containing only one player.

In order to retain only valid continuations, an audio LLM was used to filter continuations, followed by human confirmation. Appendix D.3 details the methodology.

4.3 SCORING PROTOCOL

We use text-free MusicGen-large to score continuations from all four generators, ranking candidates within a common token space. MusicGen is also one of the generators; Section 6.2 tests agreement when its outputs are excluded. We additionally score the same clips with InspireMusic and Magenta RT to examine evaluator dependence.

For every context comparison, the continuation tokens must remain fixed (Equation 1). With MusicGen, we encode the continuation once, encode each context separately, and concatenate their codes for scoring. Score differences therefore compare the predictability of exactly the same sequence. Encoding context and continuation together instead changed 19% of target tokens on average relative to encoding the continuation alone in our study set; the induced change in score was comparable to the condition contrasts (Appendix C.1).

For decoy-based measures, the common evaluator averages contrasts over eight unrelated contexts per item. Decoy assignments are shared across generators. Appendix C details context construction.

5 CONTROLLED EVALUATION OF CONTEXT USE

Do the measures distinguish changes in the context used to generate music? We designed the generation conditions to test whether the self and partner measures respond differently when the information available about each player changes. Figure 2 varies scoring contexts for a fixed continuation. Here our main comparisons retain the target player's (A) past while providing, withholding, or replacing the partner's (B) past. We vary generation conditions to produce different continuations. We compute every PMus measure using the same scoring-context comparisons based on the original A/B pasts, then compare scores across generation conditions. Section 6 tests their relationship to listener judgments. Providing the true partner produces larger changes in the partner measures than in the self measure, supporting distinct sensitivities to generation context.

5.1 Controlled Generation Conditions

Table 2 defines four generation conditions. The generation decoy $\bar{B}$ is fixed per item and drawn from another track across collections. Scoring retains its own decoy assignments. For InspireMusic, Empty also switches to text-to-music generation and onset-based cropping, so it does not isolate audio-context removal.

Table 2: **Generation conditions.** Pasts denote musical information supplied through each model's conditioning route (Section 4.1). Condition names describe inputs, not output quality.

Condition	Musical past supplied	Generated continuation
Positive	Target A and true partner B	X_{Positive}
Partial	Target A alone	X_{Partial}
Negative	Target A and unrelated partner $\bar{B}$	X_{Negative}
Empty	No audio past, $\varnothing$	X_{Empty}

Figure 3 compares two changes to generation inputs: providing the true partner and adding an unrelated partner. For each item and generator, we subtract scores only when both conditions have a screen-accepted output. We average these differences over matched generators within each item, then over items. This avoids comparing different generator sets on opposite sides of a contrast. Of 232 items, 224 support Positive–Partial and 228 support Partial–Negative; both subsets span all 44 source tracks. Appendix C.2 gives selection counts and per-generator results.

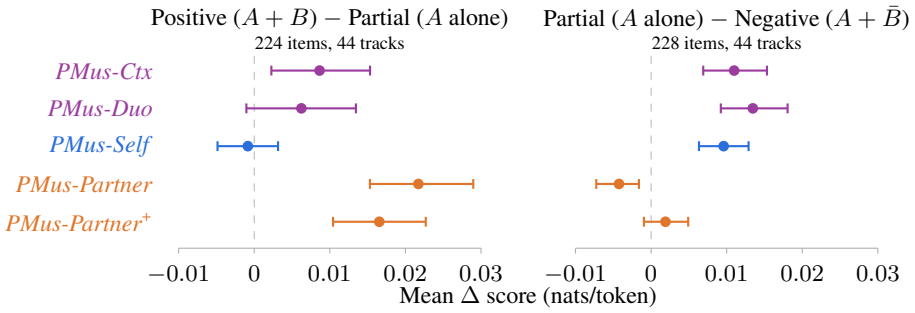


Figure 3: **Changes in *PMus* across generation conditions.** Points are mean matched differences under fixed-target MusicGen scoring. Bars are source-track bootstrap intervals (100,000 draws), Bonferroni-adjusted for approximate 95 % family-wise coverage across the ten displayed contrasts and four comparisons of *PMus-Partner* and *PMus-Partner*$^+$ with *PMus-Self* (Appendix C.2).

Different responses across measures. Providing the true partner yields substantial paired changes in *PMus-Partner* and *PMus-Partner*$^+$ ($d_z = 0.88$ and 0.80), with increases in 80–81% of matched items. Their responses also exceed *PMus-Self*'s in direct paired comparisons ($d_z = 0.79$ and 0.81), while *PMus-Self*'s own change is unresolved. Appendix C.2 gives standardized effects and their adjusted intervals; Table 8 gives raw differences. For Partial–Negative, *PMus-Self* is higher without the unrelated partner, but *PMus-Partner* is lower and *PMus-Partner*$^+$ has no resolved change. Both partner contrasts are smaller than *PMus-Self*'s (Table 8). These results establish different sensitivities to generation context.

The shared-context measures answer a broader question. *PMus-Ctx* is higher with the true partner and without an unrelated partner. *PMus-Duo* also favors the latter, while its Positive–Partial difference is unresolved after correction. These patterns are consistent with sensitivity to contextual fit, rather than identical responses from all five measures.

Source A futures can receive negative *PMus-Partner* scores, indicating that adding the partner's past can make them less predictable than under the target's past alone (Appendix C.4). A positive partner gain is therefore not a requirement for musical fit.

The Positive–Partial partner gains also appear with a common three-seed budget and with seed 0 shared across conditions (Appendix C.2). These results describe changes in predictive gain among

screen-accepted outputs. Section 6 examines their relationship to musicians' judgments of musical fit.

6 Agreement with Human Perception

6.1 Experiment: Pairwise Judgments

The controlled experiments establish different responses to generation context. We now test whether an external evaluator's predictive gains also agree with the judgments of musically trained listeners (*listeners* below), and whether the player-specific measures distinguish perceived self-continuation from partner response. After a six-musician pilot (Appendix D.2), each trial presented the same five-second past as A alone, B alone, and their mixture, followed by two five-second continuations from different generators. Each session asked either *Which track continues player A better?* or *Which track listens to player B better?*, across ten ordinary trials and two attention checks. Listeners also counted instruments (Figure 6b).

The evaluation uses human-screened continuations and judgments from eligible musicians: 208 continuation records across 46 items and 22 source tracks, yielding 940 judgments from 85 participants in 96 sessions. Appendix D.4 details eligibility, clip selection, uncertainty, and statistical controls.

6.2 Agreement with Listeners

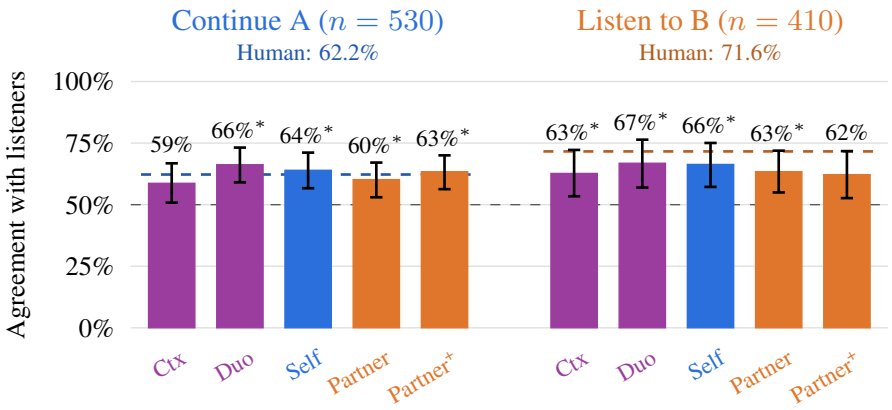


Figure 4: **Agreement between *PMus* and listeners for all five measures under both questions.** The common MusicGen evaluator ranks both continuations in each judged pair. Bars show agreement with the listener; whiskers are 95 % participant–track bootstrap intervals. Black dashes mark chance (50%); coloured dashes mark human agreement for each question. Human estimates use a smaller set of repeated pairs and are descriptive references, not ceilings. Asterisks mark above-chance agreement after nine-measure Holm correction within each question. Shared-context measures are plum, the self measure blue, and the partner measures orange, as in Figure 3. Appendix D.4 gives human-reference estimates and uncertainty; Table 16 reports the full test family.

PMus provides a signal of listener preference on these screened, cross-generator comparisons. Figure 4 shows all five measures under both questions. Under common MusicGen scoring, the shared-context measure *PMus-Duo* agrees with Continue-A and Listen-to-B choices on 66% and 67% of judgments. The designated player-specific measures follow: *PMus-Self* agrees with Continue-A choices on 64%, and *PMus-Partner* and *PMus-Partner*$^+$ with Listen-to-B choices on 63% and 62%. *PMus-Duo*, *PMus-Self*, and *PMus-Partner* remain above chance on their questions after multiplicity correction; *PMus-Partner*$^+$ does not.

Context contrasts versus direct predictability. Comparing contexts improves agreement with listeners over ranking the continuations by their predictability under the shared past alone. A direct baseline ranks clips by how predictable their audio tokens are under the evaluator, given the shared

past ($-h(X \mid C)$), without measuring what that past adds. It agrees with listeners on 41–44% of judgments. On identical pairs, PMus improves agreement by 20–23 percentage points in exploratory paired tests (Holm $p = 0.003$ for each; Appendix D.4).

Evaluation across generators and scoring models. The common MusicGen evaluator ranks outputs from all four generators without access to their probabilities. Agreement also persists when its own outputs are excluded: the remaining 270 Continue-A and 209 Listen-to-B judgments give 66% agreement for *PMus-Self*, 61% for *PMus-Partner*, and 60% for *PMus-Partner*$^{+}$. Pointwise intervals exceed chance for the first two measures; the last includes chance (Appendix D.1).

Scoring both candidates with a common InspireMusic evaluator gives 61–63% agreement for the three designated measures; Magenta RT gives 48–54%, without resolved above-chance agreement. Under correction across the full evaluator comparison, InspireMusic retains above-chance agreement for the shared-context measures and *PMus-Self* (Appendix D.1). These results support evaluation across generators and identify the scoring model as an important choice in applying PMus.

Scope of listener evidence. These results provide initial evidence linking contextual predictive gains to musicians' preferences in screened, cross-generator comparisons. *PMus-Duo* agrees with Continue-A and Listen-to-B choices on 66–67% of judgments, and the designated measures do not show a resolved advantage over this shared-context score in paired comparisons on identical judgments (Table 17); pairs on which the self and partner scores disagree show no question-specific preference either (Appendix D.4). The measures' different context comparisons provide a methodological basis for distinguishing self-continuation from partner response. Their perceptual separation remains an open hypothesis: the present study used short excerpts and screened for plausible single-player output, without selecting pairs specifically to contrast continuation and response. Differences in overall musical quality, instrumentation, or generator characteristics could have influenced both judgments, while a continuation can plausibly satisfy both questions. Listener agreement likewise did not resolve a difference between within-question and across-question comparisons (difference 0.005 $[-0.16, 0.16]$; Appendix D.4). These findings motivate a targeted experiment that varies one relationship while holding the other approximately constant. Section 7 describes this direction alongside a prospective within-generator listening study.

7 Discussion and Limitations

PMus turns questions about a continuation's relationship to its musical past into explicit comparisons of predictive gain. Individual past parts provide the reference contexts, while the continuation can remain a single audio recording. An external evaluator makes these comparisons available for human performances and outputs from open and closed source generators, without requiring the generator's probabilities. This provides a common procedure for examining shared context and individual players' contributions to the predictability of the output.

The experiments give two complementary reasons to pursue this approach. Controlled generation inputs produce different responses across the five measures, including partner gains that persist under seed-budget sensitivity checks. In the listening study, the context contrasts agree with musicians' choices more often than direct conditional likelihood on the same judgments. Together, these findings support contextual predictive gain as a useful signal for evaluating generated music. They motivate further validation of within-generator selection and of the intended self/partner interpretations.

Interpreting predictive gains. Each score describes how a chosen evaluator predicts the continuation from its past. The definition specifies an incremental predictive relationship; it does not identify causal listening inside the generator. Zero gain means equal mean surprisal under the compared contexts, with several possible explanations detailed in Appendix A. Gains can reflect both musical structure and acoustic continuity, and their alignment with listeners depends on the evaluator (Section 6.2). The present validation concerns five-second duo excerpts, with individual past parts available and judgments from musically trained listeners.

Selecting among continuations from the same generator. The listening study provides initial evidence that PMus tracks musicians' preferences across outputs from different generators. A nat-

ural next step is to test whether it also helps select among several continuations from the same generator for a given context. The present comparisons leave that application open: agreement does not clearly exceed a baseline based on average preferences between generators (Appendix D.4).

Separating self-continuation from partner response. The self and partner measures isolate different context comparisons and respond differently to generation inputs. Their distinct perceptual roles remain to be established: the listening questions do not resolve the intended distinction, including on pairs where self and partner rankings disagree (Appendix D.4). Direct comparisons also leave their ranking advantage over shared-context measures unresolved (Table 17).

The listening set was screened for plausible single-player output, rather than selected to contrast continuation and response. A targeted study could compare pairs matched on one score and contrasted on another, with selection criteria and exclusions specified in advance. A representative sample alongside these pairs would retain an estimate of agreement on ordinary examples. A prospective within-generator study would pair continuations of the same item from different seeds of one generator, so that agreement cannot be carried by generator identity. Longer excerpts could test whether the distinction becomes clearer with more musical context, while recorded performances would allow assessment separately from generated-output quality. The common scoring protocol permits these tests on both recorded and generated continuations.

Studying the influence of generation history. Beyond evaluating individual continuations, PMus offers a basis for studying how generated music depends on the history available to a generator. Future experiments could vary the duration and content of that history, removing or replacing earlier passages while keeping the recent context unchanged. Comparing the resulting continuations under a common scoring protocol would test whether changes to earlier history affect their shared and player-specific predictive relationships. Such experiments could examine whether longer available contexts translate into sustained musical relationships with earlier material.

8 Conclusion

PMus provides an information-theoretic framework for evaluating how a fixed audio continuation relates to its shared context and individual players. It measures predictive gains using the individual past parts, with no requirement to separate the output or access the generator's probabilities. Controlled experiments across four generators reveal different responses across the five measures, and the context contrasts agree with musicians on 62–67% of judgments where raw predictability does not, supporting contextual predictability as a useful signal of musical fit. These findings motivate targeted validation of continuation and response, within-generator selection, and the influence of longer generation histories. By making contextual relationships measurable under a common scoring protocol, PMus provides a practical basis for evaluating and studying generated music in relation to its past.

AI Use Statement

LLM assistants helped draft and polish text, search related work, and write analysis code; the authors verified all results and are responsible for the content. All generated audio comes from the music generators under study (Section 4.1).

Ethics Statement

Our 10-minute online listening test was taken on participants' own headphones at a moderate volume, with informed consent and pseudonymous responses. All participants who completed the task were paid, including those excluded from analysis. The study was minimal-risk and, as non-medical research, exempt from CPP review under French law. Data were handled in line with the GDPR. A similar follow-up study has been submitted for institutional ethics review.

Reproducibility Statement

Appendices B, C, and D specify corpus curation, window selection, generation settings and seeds, scoring arms, decoy assignment, exclusion rules, bootstrap seeds, and correction families. Upon acceptance we will release the scoring and analysis code, the per-arm score tables behind every reported number, the corpus and study manifests, and the listener judgments with pseudonymous participant keys. Source audio is available from URMP, Slakh, and MedleyDB under their own licences.

References

Rachel M. Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. MedleyDB: A multitrack dataset for annotation-intensive MIR research. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2014. URL https://archives.ismir.net/ismir2014/paper/000322.pdf.

Antoine Caillon, Brian McWilliams, Cassie Tarakajian, Ian Simon, Ilaria Manco, Jesse Engel, Noah Constant, Yunpeng Li, Timo I. Denk, Alberto Lalama, Andrea Agostinelli, Cheng-Zhi Anna Huang, Ethan Manilow, George Brower, Hakan Erdogan, Heidi Lei, Itai Rolnick, Ivan Grishchenko, Manu Orsini, Matej Kastelic, Mauricio Zuluaga, Mauro Verzetti, Michael Dooley, Ondrej Skopek, Rafael Ferrer, Savvas Petridis, Zalán Borsos, Äaron van den Oord, Douglas Eck, Eli Collins, Jason Baldridge, Tom Hume, Chris Donahue, Kehang Han, and Adam Roberts. Live music models. *arXiv preprint arXiv:2508.04651*, 2025. URL https://arxiv.org/abs/2508.04651.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990. URL https://aclanthology.org/J90-1003/.

Ruben Ciranni, Giorgio Mariani, Michele Mancusi, Emilian Postolache, Giorgio Fabbro, Emanuele Rodolà, and Luca Cosmo. COCOLA: Coherence-oriented contrastive learning of musical audio representations. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025. URL https://arxiv.org/abs/2404.16969.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.05284.

Shlomo Dubnov. Analysis of musical structure in audio and MIDI signals using information rate. In *Proceedings of the International Computer Music Conference*, pp. 187–192, 2006.

Shlomo Dubnov. Deep music information dynamics. *arXiv preprint arXiv:2102.01133*, 2021. URL https://arxiv.org/abs/2102.01133.

Vignesh Gokul, Chris Francis, and Shlomo Dubnov. Evaluating co-creativity using total information flow. *arXiv preprint arXiv:2402.06810*, 2024. URL https://arxiv.org/abs/2402.06810.

Junmin Gong, Yulin Song, Wenxiao Zhao, Sen Wang, Shengyuan Xu, and Jing Guo. ACE-Step 1.5: Pushing the boundaries of open-source music generation. GitHub repository, 2026. URL https://github.com/ace-step/ACE-Step-1.5.

Maarten Grachten and Javier Nistal. Accompaniment prompt adherence: A measure for evaluating music accompaniment systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. URL https://arxiv.org/abs/2503.06346.

David Huron. *Sweet Anticipation: Music and the Psychology of Expectation*. MIT Press, 2006.

Daniël Lakens. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4:863, 2013. doi: 10.3389/fpsyg.2013.00863.

Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia*, 2019. doi: 10.1109/TMM.2018.2856090. URL https://arxiv.org/abs/1612.08727.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, pp. 110–119, 2016. doi: 10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014/.

Ethan Manilow, Gordon Wichern, Prem Seetharaman, and Jonathan Le Roux. Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019. URL https://arxiv.org/abs/1909.08494.

Leonard B. Meyer. *Emotion and Meaning in Music*. University of Chicago Press, 1956.

Shinnosuke Nakayama, Vrishin R. Soman, and Maurizio Porfiri. Musical collaboration in rhythmic improvisation. *Entropy*, 22(2):233, 2020. doi: 10.3390/e22020233. URL https://www.mdpi.com/1099-4300/22/2/233.

Yatin Nandwani, Vineet Kumar, Dinesh Raghu, Sachindra Joshi, and Luis Lastras. Pointwise mutual information based metric and decoding strategy for faithful generation in document grounded dialogs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10335–10347, 2023. doi: 10.18653/v1/2023.emnlp-main.639. URL https://aclanthology.org/2023.emnlp-main.639/.

Ashwin Paranjape and Christopher Manning. Human-like informative conversations: Better acknowledgements using conditional mutual information. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 768–781, 2021. doi: 10.18653/v1/2021.naacl-main.61. URL https://aclanthology.org/2021.naacl-main.61/.

Marcus T. Pearce and Geraint A. Wiggins. Auditory expectation: The information dynamics of music perception and cognition. *Topics in Cognitive Science*, 4(4):625–652, 2012. doi: 10.1111/j.1756-8765.2012.01214.x.

Marcus T. Pearce, María Herrojo Ruiz, Selina Kapasi, Geraint A. Wiggins, and Joydeep Bhattacharya. Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage*, 50(1):302–313, 2010. doi: 10.1016/j.neuroimage.2009.12.019.

Alain Riou, Stefan Lattner, Gaëtan Hadjeres, Michael Anslow, and Geoffroy Peeters. Stem-JEPA: A joint-embedding predictive architecture for musical stem compatibility estimation. In *Proceedings of the 25th International Society for Music Information Retrieval Conference*, 2024. URL https://arxiv.org/abs/2408.02514.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025. URL https://arxiv.org/abs/2502.05139.

Ruibin Yuan, Hanfeng Lin, Shuyue Guo, Ge Zhang, Jiahao Pan, Yongyi Zang, Haohe Liu, Yiming Liang, Wenye Ma, Xingjian Du, Xeron Du, Zhen Ye, Tianyu Zheng, Zhengxuan Jiang, Yinghao Ma, Minghao Liu, Zeyue Tian, Ziya Zhou, Liumeng Xue, Xingwei Qu, Yizhi Li, Shangda Wu, Tianhao Shen, Ziyang Ma, Jun Zhan, Chunhui Wang, Yatian Wang, Xiaowei Chi, Xinyue Zhang, Zhenzhu Yang, Xiangzhou Wang, Shansong Liu, Lingrui Mei, Peng Li, Junjie Wang, Jianwei Yu, Guojian Pang, Xu Li, Zihao Wang, Xiaohuan Zhou, Lijun Yu, Emmanouil Benetos, Yong Chen, Chenghua Lin, Xie Chen, Gus Xia, Zhaoxiang Zhang, Chao Zhang, Wenhu Chen, Xinyu Zhou, Xipeng Qiu, Roger Dannenberg, Jiaheng Liu, Jian Yang, Wenhao Huang, Wei Xue, Xu Tan, and Yike Guo. YuE: Scaling open foundation models for long-form music generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a6a21421022d1da16eadbba533980530-Abstract-Conference.html.

Chong Zhang, Yukun Ma, Qian Chen, Wen Wang, Shengkui Zhao, Zexu Pan, Hao Wang, Chongjia Ni, Trung Hieu Nguyen, Kun Zhou, Yidi Jiang, Chaohong Tan, Zhifu Gao, Zhihao Du, and Bin Ma. InspireMusic: Integrating super resolution and large language model for high-fidelity long-form music generation. *arXiv preprint arXiv:2503.00084*, 2025. URL https://arxiv.org/abs/2503.00084.

A Other Measures and Interpretation

Reading incremental gains. A gain compares the predictability of the same continuation under two contexts. Noise can have high surprisal under both and therefore little gain. When the players' pasts carry similar information, adding one past can also provide little predictive improvement; identical scoring conditions give zero gain. More generally, zero means equal average surprisal and can reflect redundant pasts, evaluator insensitivity, or cancellation across tokens.

Supplementary measures. Table 3 describes three supplementary comparisons: relative prediction by the individual pasts, a normalized partner gain, and prediction relative to the source mixture future. *PMus-Attr* and *PMus-Partner*$_{\text{frac}}$ are included in the complete nine-measure listener analysis (Table 16); their specific interpretations remain exploratory.

Table 3: **Further measures considered.** Same notation as Table 1.

Measure	Definition	Reading	Relation to the main measures
PMus-Attr	$h(X \mid B) - h(X \mid A)$	positive when A alone predicts X better than B alone	equals *PMus-Self* − *PMus-Partner*; distance normalization did not establish a cross-item benefit in exploratory tests
PMus-Partner$_{\text{frac}}$	$[h(X \mid A) - h(X \mid C)]/h(X \mid A)$	partner gain as a share of the surprise under A alone	gave very similar rankings to *PMus-Partner* in exploratory tests ($r_s = 0.97$–0.98); dimensionless does not imply cross-evaluator calibration
PMus-Cal	$h(X \mid C) - h(\text{mixfut} \mid C)$	is X as predictable as what was actually played?	a typicality comparison, not an information gain; uses the recorded future as an offline reference

Relations between the measures. The NLL contrasts satisfy *PMus-Attr*=*PMus-Self*-*PMus-Partner*. *PMus-Duo* is equivalently the decoy-contrastive difference $[h(X \mid \emptyset) - h(X \mid C)] - [h(X \mid \emptyset) - h(X \mid \bar{C})]$, so full-context replacement (*PMus-Duo*) and partner replacement with A fixed (*PMus-Partner*$^{+}$) answer different questions. *PMus-Partner* and *PMus-Self* do not generally sum to *PMus-Ctx*. Mixing is not lossless joint access to the two players: a solo-to-mix contrast may also reflect masking, level changes, or a different conditioning route. The generation controls do not isolate all these effects. These model-based contrasts are not a decomposition into unique, shared, and synergistic musical information.

B Dataset and Generation Details

Source-duo curation. Curation sought two distinguishable musical lines rather than a lead melody over accompaniment, and preceded window selection and the screening of generated continuations. For Slakh, MIDI-based ranking covered all 2,100 tracks. Drum stems, pad, effect and percussive

classes were excluded; a candidate voice needed chordiness below 25% of its sounding time, activity over at least 30% of the track, at least 0.8 notes per sounding second and at least six distinct pitches. Pairs were ranked mainly by co-play, the fraction of one-second bins in which both stems are active, with small bonuses for register separation and single-note purity and a penalty for a bass voice. The top 50 tracks were auditioned as 25-second previews and rated on a 0–5 scale; ratings of at least 3 yielded 21 duos. MedleyDB selection used audio: bleed-free tracks with at least two non-vocal, non-percussion, non-bass stems, scored by co-play of RMS-active seconds; auditioning all 18 eligible tracks yielded eight duos. All 11 URMP duets were included without this screen. For MusicGen, 13 of 30 generated duos were selected by audition, then source-separated; four were retained after auditioning the separated stems. The ranking scripts, selected pairs, ratings and separation verdicts are in `data/duo_benchmark/` of the code repository.

Duos and windows. The selected URMP (Li et al., 2019), Slakh (Manilow et al., 2019), MedleyDB (Bittner et al., 2014), and MusicGen duos give 44 tracks. Nine of the 13 separated MusicGen duos were rejected for separation artefacts. Stems are resampled to 32 kHz mono and RMS-normalised; the mix is their sum, linearly peak-scaled to 0.95. Each track gives $W = \mathrm{clamp}(\mathrm{round}((L-10)/30), 1, 3)$ windows for a stem length of L seconds. Per half-second frame a stem is active when its RMS exceeds a tenth of the stem's RMS; each window's end is placed where both stems are active in the largest share of the surrounding ten seconds, searching each of W equal segments (2 of the 116 windows fall below a 0.9 share and are kept and flagged).

Waveform mixing and model conditioning. The notation $\mathrm{mix}(A, B)$ denotes waveform addition followed by linear peak scaling, as described above, rather than the unscaled sum $A + B$. This waveform operation is symmetric in A and B and defines the shared audio context used by our primary external evaluator. Generation models need not receive this mixture through a single input: their conditioning routes differ, as detailed below. For example, Magenta RT receives the target player's past as token context and the partner's past through a style embedding. Native scoring likewise follows model-specific conditioning routes (Table 5); these need not be symmetric in the two players.

Generation and clipping. MusicGen-large (Copet et al., 2023) uses the audio past as a decoder prefix and an instrument-specific text prompt, with guidance 3.0, temperature 1.0 and top-k 150. InspireMusic-1.5B (Zhang et al., 2025) uses the past as an audio prompt, empty text, the verse tag and a ten-second time request; we retain the first five seconds of its output. Magenta RT (Caillon et al., 2025) uses A's past as token context and the mean of the MusicCoCa embeddings of B's past and the text prompt as style. It generates three crossfaded two-second chunks with guidance 5.0, temperature 1.1 and top-k 40; we retain the first five seconds. Its chunk seeds are the item seed plus the chunk index. MusicGen and InspireMusic set both Torch and NumPy seeds. ACE-Step 1.5 (Gong et al., 2026) uses the turbo configuration with eight ODE steps, shift 3.0 and guidance scale 1.0 on a 30-second canvas. Its context guidance combines two velocity fields with identical initial noise, $v = v_A + s(v_{\mathrm{mix}} - v_A)$, with $s = 0.5$ for Positive and $s = 0$ for Partial. The scored continuation is seconds 5–10 of the generated canvas. Generated clips are stored as five-second, 32 kHz mono audio. YuE uses a style reference and appears only in the pilot.

Empty generation. Empty removes the audio past but retains the instrument-specific text where the generator uses it. MusicGen generates ten seconds without an audio prefix; Magenta starts with an empty token state and text-only style, generating six two-second chunks. Both retain seconds 5–10. ACE-Step switches from repainting to text-to-music on the same 30-second canvas and retains seconds 5–10. InspireMusic switches from continuation with empty text to text-to-music with the item's text prompt and a 30-second request. Its clip is seconds 5–10 after the detected onset: the first 100 ms frame whose RMS exceeds 5 % of the largest frame RMS. Empty therefore also changes task mode and, for InspireMusic, text conditioning and clipping position.

Generation decoys and seed selection. Negative retains A and replaces B with one fixed random partner from another source track. The pool spans collections; 145 of the 232 assignments cross collection boundaries. Assignments use NumPy's default random generator with seed 20260909 and are shared across seeds and generators. Controlled-condition cells retain the first Qwen-accepted clip in seed order. Positive MusicGen and InspireMusic use seeds 0–16; their Partial conditions use

0–2 and Negative and Empty have a budget of 0–9. ACE-Step uses 0–5 for Positive, Negative and Empty and 0–2 for Partial. Magenta uses 0–2 throughout. Further Negative and Empty seeds beyond 0–2 were requested for cells without an accepted clip. These are maximum budgets, not equal numbers of attempts in every cell. The selected seed can differ between conditions. Table 4 reports the resulting coverage. The listening study instead admits human-confirmed clips and can retain multiple seeds for one item and generator.

Table 4: Selected-output coverage, reported as outputs / unique source tracks. Figure 3 uses four generators; the YuE row is additional pilot/archive coverage and is excluded from that figure. Each item is a track–window–continued-player tuple, so multiple items may come from one track. GT and PGT use source futures; generated cells retain the first one-player-screened seed. Matched contrasts use subsets of these outputs.

Generator/evaluator	GT/PGT	Positive	Partial	Negative	Empty
MusicGen self	232/44	149/41	177/43	133/40	211/44
InspireMusic self	232/44	122/31	191/44	111/42	219/44
YuE self	232/44	193/43	—	—	—
Magenta RT self	232/44	217/44	167/41	216/44	168/39
ACE-Step, external MusicGen	232/44	208/44	214/44	219/44	232/44

C Scoring Implementation and Additional Robustness Tests

Scoring arms and units. An arm specifies both audio and non-audio conditioning. MusicGen's external evaluator uses null text; its own evaluator uses the item's text prompt in every arm. InspireMusic's native evaluator uses empty text, time embeddings for 0–10 seconds and the verse tag. Magenta retains the item's text in every arm. Native likelihoods are scored without sampling guidance. Each target is encoded once and its tokens reused across scoring arms; MusicGen concatenates separately encoded context and target codes. Table 5 gives the audio and style inputs. Magenta's token/style contrasts are analogous to the waveform arms rather than identical implementations. Its A-only arm drops audio style, so equal A and B waveforms need not give identical conditioning arms.

Table 5: Audio-conditioning arms. MusicGen and InspireMusic use the indicated waveform as an audio prefix. Magenta uses separate token-context and style channels. Here e_T is the item's text-style embedding and e_B the partner's audio-style embedding. Target tokens and teacher-forced target history are shared between arms.

Arm	MusicGen / InspireMusic	Magenta token past	Magenta style
Ø	none	none	e_T
A	A	A	e_T
B	B	none	$(e_T + e_B)/2$
C	$\mathrm{mix}(A, B)$	A	$(e_T + e_B)/2$
$\bar{C}$	$\mathrm{mix}(\bar{A}, \bar{B})$	$\bar{A}$	$(e_T + e_{\bar{B}})/2$
$\mathrm{mix}(A, \bar{B})$	$\mathrm{mix}(A, \bar{B})$	A	$(e_T + e_{\bar{B}})/2$

Scores average natural-log cross-entropy over valid target tokens, excluding the context. A token is a codebook token, not an audio frame. MusicGen uses four codebooks at 50 frames per second; InspireMusic uses one WavTokenizer stream at 75 tokens per second; Magenta uses 16 codebooks at 25 frames per second. Magenta scores three rolling two-second chunks, includes the fixed target history in later chunks and excludes final padding from the loss. MusicGen audio copies use linear peak scaling and -23 LUFS normalization for targets and contexts. The native InspireMusic and Magenta scorers use -23 LUFS targets and raw window contexts. Raw scores from these token spaces are not on a common scale.

Scoring decoys. Eight other-track items are assigned per target item, preferring the same collection and filling from other collections when needed. Earlier valid assignments are retained, and the generation decoy is excluded. Sampling is by item, so eight IDs need not represent eight distinct

tracks or mixed waveforms. External MusicGen uses all eight assignments for both decoy measures, with the same assignments across generators. *PMus-Duo* replaces the duo, whereas *PMus-Partner*$^{+}$ retains A and replaces B. Native Magenta uses four. Native InspireMusic uses eight. For the matched-evaluator analysis, all eight InspireMusic arms are assembled from the global NLL cache by target identity and decoy context, with text and other settings fixed. Magenta additionally uses item-specific text style to identify reusable scores.

The fixed-target runs used three RTX A6000 GPUs; checking the saved tables requires no model inference.

C.1 Fixed-Target Audit

The first version of the MusicGen scorer encoded each context and the target as one waveform. Comparing the target part of that joint encoding with the target encoded on its own, over the 300 study targets (208 continuations and 92 recorded futures; 269 distinct audio files) and 19 audio-context arms (excluding the no-context arm), 81 % of the tokens matched on average, 38 % in the first ten frames (32–49 % by target type and arm) and 91 % in the first codebook; per target, the share ranged from a median of 0.74 in its least to 0.87 in its most similar arm. The arms therefore scored different sequences. The reported joint-versus-separate NLL shift averages about 0.05 nats per token, comparable to the contrasts of Section 5.

The corrected MusicGen scorer encodes the target once, on its own. Each context is encoded separately and its codes are concatenated with the fixed target codes. Every arm therefore scores exactly the same continuation tokens. InspireMusic and Magenta already encode their targets separately. All primary results use fixed target tokens. With the target encoded once, the true context still helps the recorded futures (*PMus-Ctx* $+0.047$ nats per token for recorded A futures, $+0.177$ for recorded mixtures), showing that useful context gain survives separate encoding. This observation alone does not isolate a boundary effect. On generated study continuations, mean *PMus-Ctx* is -0.008 under separate encoding. The correction is required to compare the same event, independent of whether listener agreement improves. Table 6 compares the protocols on their shared judgment set, using item-aware score joins.

Table 6: Encoding audit on shared judgments. Both columns preserve each target's context item. Entries are listener agreement with pointwise 95 % participant–track bootstrap intervals (10,000 draws). The historical joint encoding changes target tokens across arms; separate encoding fixes them. Shared support is smaller than the final study set because the older cache lacks one clip. This descriptive comparison has no significance markers.

Measure	Question	n	Joint encoding	Separate encoding
PMus-Ctx	both	928	0.572 [0.503, 0.636]	0.602 [0.535, 0.676]
PMus-Duo	both	928	0.639 [0.581, 0.695]	0.664 [0.595, 0.730]
PMus-Self	Continue A	522	0.580 [0.499, 0.656]	0.636 [0.560, 0.709]
PMus-Partner	Listen to B	406	0.544 [0.462, 0.633]	0.633 [0.548, 0.718]
PMus-Partner^{+}	Listen to B	406	0.616 [0.526, 0.698]	0.618 [0.522, 0.714]

Boundary sensitivity. Dropping the first 0.2 s or 0.5 s of target loss gives *PMus-Self* agreement of 0.609 and 0.630 on Continue A, compared with 0.640 for the full target. *PMus-Partner* agreement on Listen to B changes from 0.634 to 0.602 and 0.595; *PMus-Partner*$^{+}$ changes from 0.622 to 0.627 and 0.620. Fixing target tokens from one encoding with the true context instead gives 0.598, 0.598, and 0.629, respectively. Positive associations remain, but their magnitude is not invariant to boundary handling. These checks use saved scores and do not isolate every effect of the codec splice.

C.2 Detailed Controlled-Condition Results

The matched analysis compares conditions within each item and generator, then averages the paired differences within an item. This holds generator composition fixed across each contrast. Table 7 reports all five measures for the four generators, with pointwise source-track bootstrap intervals (4,000 draws). Comparisons describe screen-accepted outputs; seed-budget sensitivity is examined below.

Table 7: All matched-generator contrasts under external MusicGen. Entries are mean differences in nats per token, with unadjusted 95 % source-track bootstrap intervals (4,000 draws). Coverage is paired items/source tracks for each contrast. The matched pooled block averages within-generator differences within each item before pooling items. Results describe screen-accepted outputs, with seeds selected separately in each condition.

Measure	Positive–Partial	Partial–Negative
MusicGen: Positive–Partial 122/35; Partial–Negative 115/38		
PMus-Ctx	$+0.0295$ [+0.0200, +0.0405]	$+0.0136$ [+0.0089, +0.0188]
PMus-Duo	$+0.0285$ [+0.0193, +0.0383]	$+0.0136$ [+0.0085, +0.0190]
PMus-Self	$+0.0017$ [-0.0045, +0.0090]	$+0.0082$ [+0.0043, +0.0123]
PMus-Partner	$+0.0496$ [+0.0420, +0.0580]	-0.0078 [-0.0127, -0.0032]
PMus-Partner^{+}	$+0.0386$ [+0.0308, +0.0476]	$+0.0016$ [-0.0022, +0.0050]
InspireMusic: Positive–Partial 113/30; Partial–Negative 94/39		
PMus-Ctx	$+0.0355$ [+0.0248, +0.0463]	$+0.0329$ [+0.0256, +0.0408]
PMus-Duo	$+0.0324$ [+0.0219, +0.0425]	$+0.0363$ [+0.0279, +0.0450]
PMus-Self	$+0.0049$ [-0.0023, +0.0119]	$+0.0302$ [+0.0233, +0.0371]
PMus-Partner	$+0.0666$ [+0.0545, +0.0779]	-0.0144 [-0.0209, -0.0083]
PMus-Partner^{+}	$+0.0542$ [+0.0454, +0.0629]	$+0.0061$ [+0.0001, +0.0123]
Magenta RT: Positive–Partial 158/40; Partial–Negative 162/40		
PMus-Ctx	-0.0033 [-0.0067, +0.0001]	$+0.0038$ [+0.0003, +0.0072]
PMus-Duo	-0.0029 [-0.0061, +0.0001]	$+0.0043$ [+0.0012, +0.0075]
PMus-Self	$+0.0005$ [-0.0028, +0.0037]	$+0.0015$ [-0.0007, +0.0037]
PMus-Partner	-0.0005 [-0.0037, +0.0027]	$+0.0014$ [-0.0016, +0.0045]
PMus-Partner^{+}	-0.0001 [-0.0024, +0.0022]	$+0.0013$ [-0.0013, +0.0039]
ACE-Step: Positive–Partial 202/44; Partial–Negative 209/44		
PMus-Ctx	$+0.0011$ [-0.0024, +0.0045]	$+0.0066$ [+0.0030, +0.0104]
PMus-Duo	-0.0041 [-0.0073, -0.0007]	$+0.0116$ [+0.0080, +0.0156]
PMus-Self	-0.0044 [-0.0076, -0.0012]	$+0.0085$ [+0.0053, +0.0120]
PMus-Partner	$+0.0084$ [+0.0056, +0.0113]	-0.0030 [-0.0052, -0.0005]
PMus-Partner^{+}	$+0.0056$ [+0.0033, +0.0081]	$+0.0008$ [-0.0014, +0.0030]
Matched pooled: Positive–Partial 224/44; Partial–Negative 228/44		
PMus-Ctx	$+0.0086$ [+0.0043, +0.0132]	$+0.0110$ [+0.0082, +0.0139]
PMus-Duo	$+0.0062$ [+0.0013, +0.0113]	$+0.0135$ [+0.0107, +0.0167]
PMus-Self	-0.0008 [-0.0036, +0.0019]	$+0.0096$ [+0.0074, +0.0119]
PMus-Partner	$+0.0217$ [+0.0171, +0.0266]	-0.0042 [-0.0061, -0.0024]
PMus-Partner^{+}	$+0.0165$ [+0.0124, +0.0208]	$+0.0019$ [+0.0000, +0.0040]

Coverage and adjusted comparisons. An item is a source window with one stem designated as A: 116 windows in 44 tracks give 232 items. An item enters a contrast when at least one generator has a selected, scored output in both conditions. Positive–Partial retains 224 items and 595 matched item/generator pairs. Eight items have outputs in both conditions but no generator in common. Partial–Negative retains 228 items and 580 matched pairs; three items have disjoint generator sets and one has no selected Negative output. Both retained sets cover all 44 tracks. Items receive equal weight after averaging their available matched-generator differences; tracks with more items contribute more to the point estimate.

Figure 3 and Table 8 use 100,000 whole-source-track bootstrap draws (seed 20260924). Each draw retains a track's windows and both player directions together. We apply Bonferroni adjustment across 14 contrasts: five measures under two condition comparisons, plus *PMus-Partner* minus *PMus-Self* and *PMus-Partner*$^+$ minus *PMus-Self* under each comparison. Each percentile interval uses tail probabilities $0.05/(2 \cdot 14)$, targeting approximate 95 % family-wise coverage. Between-measure differences use the same matched item/generator pairs, so their uncertainty preserves covariance between measures. This exploratory analysis uses its own 14-contrast correction family. The advance expectations are recorded in Appendix C.3; Table 7 reports pointwise per-generator intervals.

Table 8: Matched condition differences and direct between-measure comparisons. Entries are means and Bonferroni-adjusted source-track bootstrap intervals for the 14-contrast family (100,000 draws; approximate 95 % family-wise coverage). The last two rows compare condition differences, not absolute scores.

Measure or comparison	Positive–Partial	Partial–Negative
PMus-Ctx	$+0.0086$ $[+0.0023, +0.0153]$	$+0.0110$ $[+0.0069, +0.0153]$
PMus-Duo	$+0.0062$ $[-0.0010, +0.0134]$	$+0.0135$ $[+0.0092, +0.0181]$
PMus-Self	-0.0008 $[-0.0049, +0.0032]$	$+0.0096$ $[+0.0063, +0.0129]$
PMus-Partner	$+0.0217$ $[+0.0153, +0.0290]$	-0.0042 $[-0.0073, -0.0016]$
PMus-Partner$^+$	$+0.0165$ $[+0.0104, +0.0227]$	$+0.0019$ $[-0.0009, +0.0049]$
PMus-Partner $-$ *PMus-Self*	$+0.0226$ $[+0.0170, +0.0285]$	-0.0138 $[-0.0185, -0.0094]$
PMus-Partner$^+$ $-$ *PMus-Self*	$+0.0174$ $[+0.0128, +0.0218]$	-0.0077 $[-0.0105, -0.0045]$

Standardized paired effects. To express each change relative to its variability, we report paired $d_z = \overline{\delta}/s_\delta$ (Lakens, 2013). Here δ_i is an item's condition difference, averaged over its matched generators, and s_δ is the sample standard deviation across items. For between-measure comparisons, δ_i is the difference between their condition responses on the same item, not the difference between two separately standardized effects. Table 9 uses the same 224 or 228 items as Figure 3.

We recompute both the mean and standard deviation in each of 100,000 whole-track bootstrap draws (seed 20260924). Percentile intervals use Bonferroni tails $0.05/(2 \cdot 14)$ across the 14 standardized effects. Coverage is approximately 95 % within this table's family. Positive fractions describe the share of items with $\delta_i > 0$.

For Positive–Partial, *PMus-Partner* and *PMus-Partner*$^+$ increase in 180/224 and 181/224 items. Their excess responses over *PMus-Self* have $d_z = 0.79$ $[0.59, 1.03]$ and 0.81 $[0.55, 1.09]$. These effect sizes describe screened outputs under MusicGen scoring, standardized by the paired-difference variability after matched-generator averaging.

Descriptive pooled summaries. Table 10 additionally reports medians and pooled item-paired differences, averaging available generators in each condition. Generator composition differs in 154/232 Positive–Partial and 172/231 Partial–Negative pairs; selected seeds can also differ. These summaries provide the reference distributions for Appendix C.4 and the comparison with advance expectations in Appendix C.3. Inference above uses matched generators and source-track resampling.

Screened budget and shared-seed sensitivities. We restrict the saved first-accepted outputs either to seeds 0–2, imposing a common maximum three-seed budget, or to seed 0 in both conditions.

Table 9: Standardized paired effects under the common MusicGen evaluator. Entries give d_z and adjusted whole-track bootstrap intervals. The 14 intervals target approximate 95 % family-wise coverage within this table. Positive (%) is the percentage of item-level differences above zero, not classification accuracy. The final two rows compare condition responses between measures.

Measure or comparison	Positive–Partial		Partial–Negative	
	d_z [interval]	Positive (%)	d_z [interval]	Positive (%)
PMus-Ctx	$+0.37$ $[+0.10, +0.65]$	64.7	$+0.54$ $[+0.33, +0.80]$	74.1
PMus-Duo	$+0.26$ $[-0.05, +0.55]$	56.7	$+0.68$ $[+0.48, +0.93]$	75.4
PMus-Self	-0.05 $[-0.28, +0.17]$	45.1	$+0.54$ $[+0.33, +0.76]$	71.1
PMus-Partner	$+0.88$ $[+0.67, +1.14]$	80.4	-0.25 $[-0.41, -0.10]$	36.0
PMus-Partner$^+$	$+0.80$ $[+0.52, +1.12]$	80.8	$+0.14$ $[-0.08, +0.37]$	56.6
PMus-Partner $-$ *PMus-Self*	$+0.79$ $[+0.59, +1.03]$	82.1	-0.53 $[-0.72, -0.35]$	26.8
PMus-Partner$^+$ $-$ *PMus-Self*	$+0.81$ $[+0.55, +1.09]$	81.7	-0.42 $[-0.62, -0.24]$	32.0

(a) median per condition, nats/token (GT, PGT: source A+B and A futures)

measure	GT	PGT	Pos	Part	Neg	Emp
PMus-Ctx	0.186	0.047	−0.021	−0.025	−0.038	−0.073
PMus-Duo	0.274	0.138	0.036	0.037	0.017	0.002
PMus-Self	0.119	0.112	0.022	0.029	0.014	0.001
PMus-Partner	0.119	−0.065	0.014	−0.006	−0.001	0.001
PMus-Partner$^+$	0.158	0.015	0.015	0.003	0.000	−0.001

(b) descriptive mean within-item differences; available generators may differ between conditions

measure	Pos−Part	Part−Neg	Pos−Neg	Pos−Emp	GT−PGT	Part−Emp	Neg−Emp
PMus-Ctx	0.005	0.014	0.019	0.053	0.143	0.048	0.034
PMus-Duo	0.000	0.020	0.020	0.036	0.149	0.037	0.016
PMus-Self	−0.006	0.015	0.009	0.024	0.006	0.030	0.015
PMus-Partner	0.022	−0.006	0.016	0.015	0.217	−0.008	−0.002
PMus-Partner$^+$	0.015	0.003	0.018	0.019	0.165	0.004	0.001

Table 10: Descriptive condition results under fixed-target external MusicGen scoring. GT and PGT denote source mixture and source A futures. Each item/condition averages available screened generators, which can differ across conditions. No uncertainty or significance is encoded; matched contrasts and track intervals appear in Table 7.

Table 11 retains all five measures under both contrasts. Matching and item averaging follow the primary analysis; pointwise intervals use 10,000 whole-track bootstrap draws (seed 20260924). The Positive–Partial partner gains persist under both restrictions, while *PMus-Self* remains unresolved. For Partial–Negative, *PMus-Partner*'s negative change under the common budget is approximately zero with shared seed 0. These comparisons retain screen-accepted outputs; using the same seed does not hold the generated continuation fixed.

Table 11: Screened budget and shared-seed sensitivities. Entries are mean differences in nats/token with pointwise 95 % source-track bootstrap intervals (10,000 draws). Under the three-seed budget, Positive–Partial uses 498 matched cells / 222 items / 44 tracks and Partial–Negative uses 495 / 224 / 44. With seed 0 in both conditions, support is 243 / 157 / 40 and 238 / 154 / 42, respectively.

Contrast	Measure	First accepted seed ≤ 2	Seed 0 in both
Positive–Partial	*PMus-Ctx*	$+0.0052\,[+0.0009, +0.0096]$	$+0.0070\,[+0.0008, +0.0143]$
	PMus-Duo	$+0.0031\,[-0.0015, +0.0079]$	$+0.0061\,[+0.0005, +0.0123]$
	PMus-Self	$-0.0013\,[-0.0040, +0.0014]$	$+0.0009\,[-0.0025, +0.0046]$
	PMus-Partner	$+0.0153\,[+0.0107, +0.0202]$	$+0.0146\,[+0.0084, +0.0209]$
	PMus-Partner$^+$	$+0.0122\,[+0.0081, +0.0164]$	$+0.0130\,[+0.0076, +0.0186]$
Partial–Negative	*PMus-Ctx*	$+0.0092\,[+0.0062, +0.0123]$	$+0.0062\,[+0.0026, +0.0099]$
	PMus-Duo	$+0.0120\,[+0.0089, +0.0151]$	$+0.0076\,[+0.0038, +0.0117]$
	PMus-Self	$+0.0084\,[+0.0060, +0.0108]$	$+0.0058\,[+0.0031, +0.0086]$
	PMus-Partner	$-0.0029\,[-0.0049, -0.0010]$	$+0.0001\,[-0.0025, +0.0028]$
	PMus-Partner$^+$	$+0.0014\,[-0.0005, +0.0034]$	$+0.0013\,[-0.0009, +0.0038]$

C.3 Advance Expectations for the Controlled Conditions

Table 12 compares the expected orderings from a dated internal analysis plan with the descriptive pooled outcomes. The plan was not independently registered. *PMus-Duo* was initially grouped with the partner measures; its small Positive–Partial difference and larger Partial–Negative difference motivated the subsequent shared-context interpretation. The matched analysis in Table 8 gives a positive Positive–Partial point estimate for *PMus-Duo*, with an adjusted interval including zero.

Table 12: Advance expectations and unmatched pooled outcomes. Expected $\approx$ denotes a hypothesized small difference, not a demonstrated equivalence. Pooled patterns are descriptive because screened generator composition varies by condition; they are not the matched contrasts prioritized in Figure 3.

Measures	Expected	Pooled outcome
PMus-Partner, *PMus-Partner*$^+$	GT $\geq$ PGT $>$ Pos $\geq$ Part $\approx$ Neg $\approx$ Emp ≈ 0	GT exceeds PGT; Pos exceeds Part. PGT falls below Pos for *PMus-Partner* and has the same rounded median for *PMus-Partner*$^+$. Part and Neg are not identical.
PMus-Duo (initially a partner measure)	same as above	Pos and Part have close medians; both exceed Neg.
PMus-Self	PGT high; Pos $\approx$ Part $\approx$ Neg > 0; Emp ≈ 0	PGT high and Emp near zero, but the other conditions differ: Part exceeds Pos, which exceeds Neg.
PMus-Ctx	GT $\geq$ PGT $\geq$ Pos $\approx$ Part $\approx$ Neg $>$ Emp ≈ 0	GT and PGT exceed generated conditions. All generated medians are negative, including Emp (-0.073), contrary to the sign expectations.

C.4 Source Futures and the Empty Condition

Source futures show how the measures score continuations of the original passage. Figure 5 compares the source mixture future (GT) and A's future alone (PGT) with generated continuations. The corpus includes recorded performances, synthesized stems, and separated generated audio. These are reference continuations, rather than ground truth for a player's response; their score ordering is empirical.

The median *PMus-Partner* for source A futures is -0.065 nats per token (Table 10): for at least half these continuations, adding B's past reduces predictability relative to A's past alone. Positive partner gain is therefore not a requirement for musical fit. The definition also allows positive gains for solo continuations when the shared past helps predict them.

The figure additionally includes Empty, with pooled medians in Table 10. The advance plan included Positive–Empty comparisons; here we report Empty descriptively. For InspireMusic, Empty changes task mode, text conditioning, and cropping as well as audio context (Appendix B).

Under the common text-free evaluator, exchanging A and B for the same mixture future preserves the mixed context and exchanges *PMus-Self* with *PMus-Partner*, so their pooled GT distributions coincide when both directions of every item are retained.

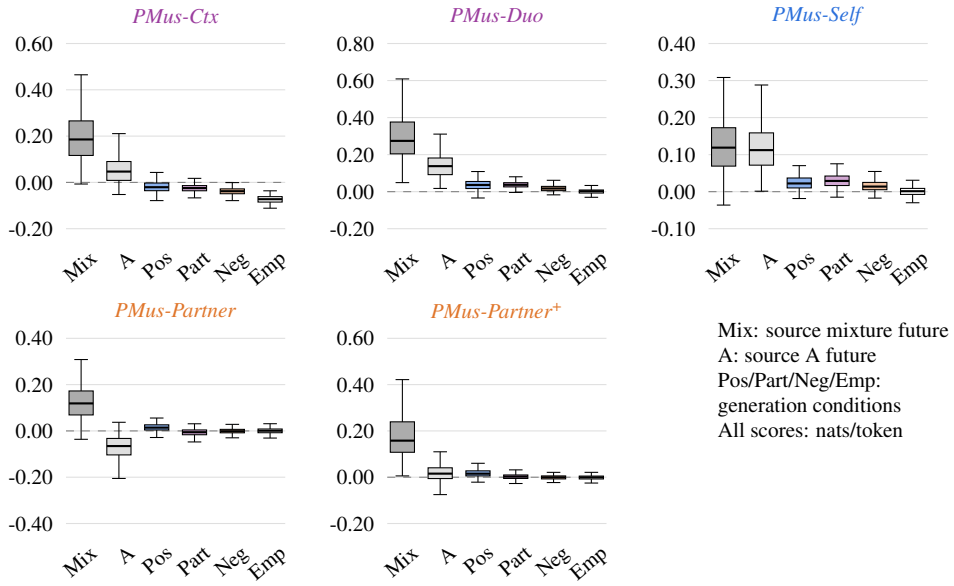


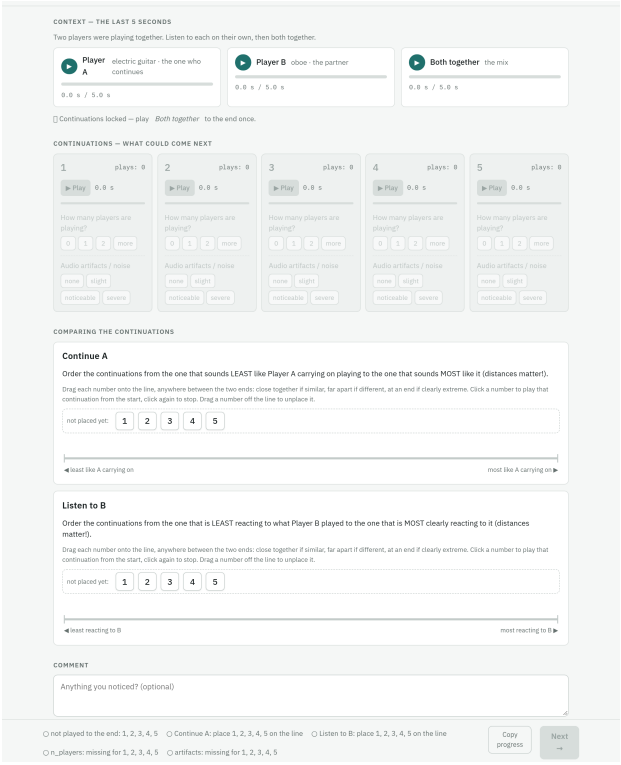
Figure 5: Source futures alongside generated conditions under the external evaluator. Mix (GT) contains both source futures; A (PGT) contains A alone. The pooled GT *PMus-Self* and *PMus-Partner* distributions coincide through direction-swap symmetry. Generated-condition item means use available screened generators, whose composition can differ by condition. Boxes span the interquartile range, with median lines and whiskers to the most extreme values within 1.5 times that range. Outliers are omitted; boxes describe the score distributions.

D Human Study Materials and Statistical Analysis

D.1 Matched Common Evaluators

One common evaluator at a time. The saved evaluator matrix lets us score both candidates of every judged pair with the same model. This avoids mixing native score scales within a pair. Table 13 applies MusicGen, InspireMusic, and Magenta to the same selected clips. Target/context identity is preserved in each join. Intervals resample the 85 participants and 22 source tracks independently, with product weights on judgments (10,000 draws). Exact score ties are omitted: Magenta *PMus-*

(a) Pilot: one trial, five continuations, two orderings



(b) Prolific study: one trial, two continuations, one question

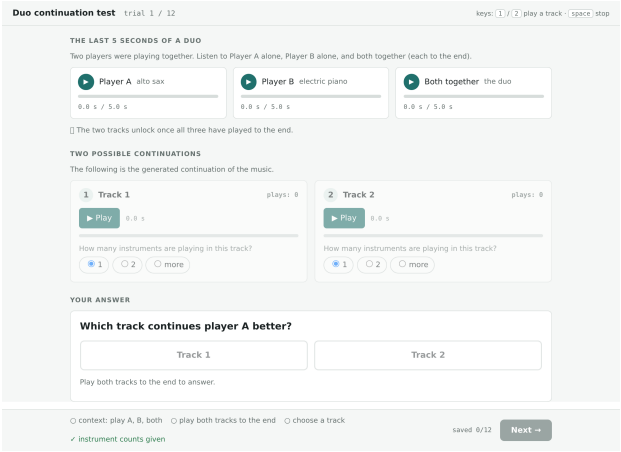


Figure 6: One trial of each listening interface, as the participants saw it. **(a)** The pilot (Appendix D.2): the duo context on top (player A alone, player B alone, both together, each played to the end before the continuations unlock), five generated continuations with a player count and an artifact rating each, and the two ordering lines — one for *continue A*, one for *listen to B* — on which the five numbers are dragged, distances included. **(b)** The Prolific study (Section 6.1): the same context block, then two continuations of the same item with a player count each, and a single forced choice — here *which track continues player A better?* — which unlocks once both tracks have played to the end. Each session shows 12 such trials, two of them attention checks.

Partner retains 397 of 410 Listen-to-B judgments; all other cells have full coverage. Paired evaluator comparisons use the common non-tied subset.

Table 13: Exploratory matched-evaluator agreement, with pointwise 95 % participant–track bootstrap intervals. Each column uses one evaluator for all generators, not a mixture of own-model scales. *PMus-Ctx* and *PMus-Duo* pool both questions ($n = 940$); *PMus-Self* uses Continue A ($n = 530$); partner measures use Listen to B ($n = 410$, except Magenta *PMus-Partner*: 397). Asterisks mark $p < 0.05$ after Holm correction over all 15 cells.

Measure	MusicGen	InspireMusic	Magenta RT
PMus-Ctx	0.604^* [0.537, 0.678]	0.627^* [0.551, 0.694]	0.510 [0.439, 0.577]
PMus-Duo	0.665^* [0.597, 0.731]	0.652^* [0.578, 0.717]	0.541 [0.478, 0.602]
PMus-Self	0.640^* [0.567, 0.711]	0.619^* [0.545, 0.692]	0.475 [0.391, 0.562]
PMus-Partner	0.634^* [0.550, 0.719]	0.610 [0.511, 0.707]	0.511 [0.415, 0.593]
PMus-Partner$^+$	0.622 [0.527, 0.717]	0.627 [0.535, 0.717]	0.541 [0.433, 0.644]

InspireMusic retains above-chance shared-context and *PMus-Self* agreement under this correction. Magenta does not resolve above-chance agreement in any cell. These results identify evaluator choice as a determinant of listener alignment across the scoring implementations tested here.

Excluding the evaluator's own generator. As an exploratory sensitivity check, we exclude every pair containing a MusicGen-generated candidate while retaining MusicGen as evaluator. This leaves 270 Continue-A judgments from 53 people and 209 Listen-to-B judgments from 39 people. Both subsets cover all 46 items and 22 tracks. Agreement is 0.659 [0.559, 0.754] for *PMus-Self*, 0.612 [0.508, 0.714] for *PMus-Partner*, and 0.603 [0.483, 0.726] for *PMus-Partner*$^+$. Pointwise intervals use the same 10,000 full-study participant–track resamples, restricted to the retained judgments; no tests are added. The *PMus-Self* and *PMus-Partner* pointwise intervals remain above chance without comparisons to the evaluator's own generated audio; the *PMus-Partner*$^+$ interval includes chance. These are still cross-generator comparisons.

D.2 Pilot Listening Experiment

Six musicians (3–29 years of reported musical experience) each completed ten trials; Figure 6a shows one of them as the listeners saw it. The 60 trials used 60 unique items drawn from 38 source tracks; each trial presented a five-second context as A alone, B alone, and the A+B mixture, followed by five continuations of A, each five seconds long. The continuations came from MusicGen, Magenta RT, InspireMusic, YuE, and ACE-Step and were shown in counterbalanced order with model identity hidden. This yielded 300 rated clips. For every trial, the listener placed the five continuations on two separate lines: from least to most like A continuing, and from least to most clearly reacting to B. They also reported the perceived number of players and artifact severity for every continuation.

The two human orderings were 70.8% concordant within trials ($\tau = 0.42$). Concordance averages agreement over the ten clip pairs in each trial, omitting tied human rankings and weighting trials equally. Each item was rated by one musician; the main study supplies the listener–listener reference and the primary measure-agreement estimates. The broader screening archive includes ten further trials from a seventh listener, giving 350 rated clips for the automatic-judge validation below.

D.3 One-Player Screening and Validation

Setup. Each judge received the five-second continuation alone, without its context, with the system prompt "You are an expert music listener. You answer questions about short audio clips precisely and only in the requested format." and the question "How many distinct players (separate instrumental or vocal parts) are playing in this clip? Answer with exactly one of: 0, 1, 2, more. Answer with that token only." Decoding was greedy, with one retry appended with "Reply with the answer only." if the reply could not be parsed; no reply of Qwen3-Omni needed the retry. The models were `Qwen/Qwen2.5-Omni-7B`, `Qwen/Qwen3-Omni-30B-A3B-Instruct`, and `nvidia/audio-flamingo-next-hf`. The reference labels are the seven-listener screening archive's answers to the same question (one listener per clip); 322 of the 350 rated clips could be

matched to a judged file, 112 of them heard as one player. Source A futures (70 clips) served as hidden single-player controls.

Agreement with the listeners. Table 14 reports four-way agreement and one-player detection (listener "1" against judge "1"). Qwen3-Omni has balanced accuracy 0.69. The listening study uses the judge to prioritize clips for human confirmation. Restricting the reference to the six musicians gives a similar balanced accuracy, 0.686.

Table 14: Automatic judges against the screening archive's player count ($n = 322$ clips). The detector columns treat "one player" as the positive class; κ is binary agreement. Controls: share of the 70 recorded single-player futures answered "1".

Judge	4-way	Sens.	Spec.	Bal. acc.	Prec.	κ	Controls
Qwen3-Omni-30B	0.525	0.62	0.76	0.69	0.58	0.37	0.90
Qwen2.5-Omni-7B	0.460	0.91	0.34	0.63	0.43	0.20	0.96
Audio Flamingo Next	0.143	0.00	1.00	0.50	0.00	−0.01	0.00

Corpus screen and author tagging. The corpus screen keeps a clip when the greedy first token is "1". Over the continuations generated from the full A+B past: MusicGen 691 kept of 3944 (5.7 attempts per kept clip over 17 seeds), InspireMusic 904 of 3944, ACE-Step 617 of 1392, YuE 338 of 695, Magenta RT 557 of 696; recorded A futures 91 % kept, recorded mixtures 52 %. The author then tagged judge-accepted clips with the listeners' question and without seeing the generator or the judge's answer; 600 clips were tagged (326 "1", 241 "2", 21 "0", 12 "more"). The five recorded A futures reached among hidden controls were all tagged "1". The dataset rule (Section 4.2) then takes every clip that a human heard as one player, 440 among the confirmed continuations, and the study uses the 46 items for which all four generators have a confirmed clip.

D.4 Listening Study Design and Analysis

Recruitment, study information, and payment. Recruitment in September 2026 targeted adults on Prolific with at least 95 % approval, English fluency, desktop audio access, an interest in music, and at least three years of instrumental experience. The stated reward rate was £9.60 per hour. Before accepting, participants saw a task description, an expected duration of about ten minutes, headphone and quiet-room requirements, notice of checks, and a contact address. Participation was voluntary; participants could return their submission without an approval-rate penalty. The welcome screen records consent through a required agreement box before trials.

Training was reassessed within the study because platform screening is self-reported. Initially, submissions reporting less than three years were excluded after completion and returned unpaid; these participants have since been paid in full. A later version screened such participants out before trials with a small fixed payment.

Participant and session accounting. Table 15 summarizes session eligibility and attention-check versions. The 96 eligible sessions came from 85 participants: 74 completed one session and 11 completed two. Continue A has 54 sessions from 53 people; Listen to B has 42 sessions from 39 people. Seven people answered both questions. There are 530 Continue-A and 410 Listen-to-B judgments; twenty timed-out trials account for the difference from 960 scheduled ordinary trials. Median reported training was seven years; headphones were reported in 95 of 96 sessions. The separate pre-trial screenout export contains 15 records. Test sessions are excluded.

The earlier attention-check protocol specifies A's future as the Continue-A target and B's as the Listen-to-B target. Its records retain distractor identities and correct sides, but not both presented audio paths. Analysis uses the stored correct answer for each session's task version.

Selected clips and coverage. The listening manifest contains 208 clip records from 46 items and 22 source tracks, including 195 distinct audio files by content hash. Multiple seeds can be retained for an item and generator. Of the 208 records, 32 retain a one-player label from the pilot screening pool and 176 retain an author's label. They represent 31 and 164 distinct audio files, respectively. Counts use the final screening-label source, with author labels taking precedence when both

Table 15: Session eligibility and attention-check versions. Exclusions are sequential; every retained session passed both checks.

Stage or version	Sessions	Details
Recorded	138	Before eligibility exclusions
Incomplete or invalid	7	Excluded
Failed attention check	7	Excluded from 131 complete valid sessions
Less than three years' training	28	Excluded after attention checks
Eligible	96	85 participants; 940 ordinary judgments
Pink-noise checks	93	Source A future versus the same future with pink noise; music RMS 6 dB above noise RMS; 186 checks
Earlier instrument checks	3	Another recorded instrument as distractor; one Continue-A and two Listen-to-B sessions; six checks

are available. The final manifest and `build_dataset.py` document this rule. The 359 cross-generator pairs were scheduled under both questions. Among eligible responses, 355 pairs have a Continue A judgment and 341 have a Listen to B judgment; 337 have both.

Measure agreement and uncertainty. A measure agrees with a judgment when it gives the chosen clip a higher score. Exact ties and missing scores are omitted for that measure. Chance is 0.5. Contextual scores are joined by item, generator, and seed. Context-free Audiobox scores (Tjandra et al., 2025) can be joined by audio identity; their available judgments are not restricted by native-evaluator coverage.

Primary intervals use 10,000 crossed bootstrap draws with seed 20260924. Each draw samples the 85 participants and 22 source tracks independently with replacement. A judgment's weight is the product of its participant's and track's sampling counts. Both sessions from a repeated participant receive the same participant weight. We report percentile 95 % intervals and two-sided bootstrap tail areas around 0.5. Holm correction retains the original nine measures per question: *PMus-Ctx*, *PMus-Duo*, *PMus-Self*, *PMus-Partner*, *PMus-Partner*$^{+}$, *PMus-Attr*, *PMus-Partner*$_{\text{frac}}$, content enjoyment, and production quality. The matched-evaluator table uses a separate exploratory family of 15 cells. These pointwise intervals and corrected tail areas answer different questions; an interval excluding chance need not pass the broader correction. In paired comparisons below, intervals spanning zero leave the difference unresolved and do not establish equivalence.

The full question-by-measure results make the scope of the primary tests explicit. Table 16 reports all nine measures under both questions; CE and PQ denote Audiobox content enjoyment and production quality. The designated measures in Section 6.2 follow their intended question mappings.

Player-specific versus shared-context agreement. We compare each designated player-specific measure with both shared-context measures on its intended listening question, using identical judgments. These six exploratory paired comparisons use 10,000 participant–track bootstrap draws, with pointwise 95 % intervals and Holm correction across the six tests. Table 17 reports the differences in agreement; the intervals leave their relative ranking performance unresolved.

Between-measure and listener–listener agreement. On the judged pairs, agreement between measure rankings ranges from 0.68 to 0.85 under the common evaluator. Differences in listener agreement use shared judgments and the same bootstrap weights. Pooling both questions, *PMus-Duo* minus *PMus-Ctx* agreement is $+0.061$ [-0.003, 0.128]. On Listen to B, *PMus-Partner*$^{+}$ minus *PMus-Partner* is -0.012 [-0.092, 0.070]. Listener–listener comparisons exclude self-comparisons and weight both participants and their shared source track. This leaves 511 same-question and 596 cross-question comparisons. The cross-minus-same agreement difference is 0.005 [-0.162, 0.155]. Within-question listener agreement is 0.622 [0.483, 0.767] for Continue A and 0.716 [0.417, 1.000] for Listen to B. These references use repeated pairs, whereas measure agreement uses all judged pairs, so they are not agreement ceilings.

Table 16: Complete nine-measure family by question. Accuracy is judgment-level agreement with pointwise 95% participant × source-track bootstrap intervals (10,000 draws). Bootstrap tail areas for accuracy versus 0.5 are Holm-adjusted within each nine-measure question family. Above-chance agreement and direct comparisons between measures are separate tests.

Question	Measure	n	Accuracy [95% CI]	$p_{\mathrm{Holm},9}$
Continue A	*PMus-Ctx*	530	$0.587\ [0.508, 0.668]$	0.1580
Continue A	*PMus-Duo*	530	$0.662\ [0.591, 0.732]$	0.0018
Continue A	*PMus-Self*	530	$0.640\ [0.567, 0.711]$	0.0028
Continue A	*PMus-Partner*	530	$0.602\ [0.530, 0.671]$	0.0480
Continue A	*PMus-Partner*$^{+}$	530	$0.634\ [0.563, 0.700]$	0.0018
Continue A	*PMus-Attr*	530	$0.566\ [0.482, 0.656]$	0.3396
Continue A	*PMus-Partner*$_{\mathrm{frac}}$	530	$0.572\ [0.495, 0.645]$	0.2720
Continue A	CE	522	$0.579\ [0.481, 0.685]$	0.3396
Continue A	PQ	522	$0.519\ [0.424, 0.610]$	0.6806
Listen to B	*PMus-Ctx*	410	$0.627\ [0.534, 0.722]$	0.0432
Listen to B	*PMus-Duo*	410	$0.668\ [0.570, 0.764]$	0.0128
Listen to B	*PMus-Self*	410	$0.663\ [0.572, 0.751]$	0.0054
Listen to B	*PMus-Partner*	410	$0.634\ [0.550, 0.719]$	0.0182
Listen to B	*PMus-Partner*$^{+}$	410	$0.622\ [0.527, 0.717]$	0.0536
Listen to B	*PMus-Attr*	410	$0.522\ [0.429, 0.618]$	1.0000
Listen to B	*PMus-Partner*$_{\mathrm{frac}}$	410	$0.634\ [0.537, 0.731]$	0.0432
Listen to B	CE	406	$0.606\ [0.515, 0.698]$	0.0858
Listen to B	PQ	406	$0.532\ [0.438, 0.615]$	1.0000

Table 17: Exploratory direct accuracy contrasts on identical judgments. The six comparisons were fixed before this additional calculation. Differences and pointwise 95% crossed-bootstrap intervals are in percentage points; tail areas are Holm-adjusted across all six comparisons.

Question	Designated minus shared	n	Difference [95% CI]	$p_{\mathrm{Holm},6}$
Continue A	*PMus-Self* − *PMus-Ctx*	530	$+5.28\ [-3.18, +14.26]$	1.000
Continue A	*PMus-Self* − *PMus-Duo*	530	$-2.26\ [-8.20, +3.54]$	1.000
Listen to B	*PMus-Partner* − *PMus-Ctx*	410	$+0.73\ [-8.42, +9.31]$	1.000
Listen to B	*PMus-Partner* − *PMus-Duo*	410	$-3.41\ [-12.48, +5.25]$	1.000
Listen to B	*PMus-Partner*$^{+}$ − *PMus-Ctx*	410	$-0.49\ [-9.29, +8.24]$	1.000
Listen to B	*PMus-Partner*$^{+}$ − *PMus-Duo*	410	$-4.63\ [-11.41, +1.85]$	1.000

When measures disagree. We retain judgments for which the item-aware external *PMus-Self* and partner score differences have opposite signs. Exact ties are excluded. Table 18 reports the proportion choosing *PMus-Self*'s preferred clip under each question. A positive between-question difference would support a relative preference for *PMus-Self* on Continue A. Shared participant–track bootstrap weights preserve dependencies across questions, including people who completed both tasks. Intervals are exploratory, pointwise 95 % intervals from 10,000 draws (seed 20260924).

Table 18: Listener choices when external *PMus-Self* and a partner measure disagree. Entries give the number of judgments and the fraction choosing *PMus-Self*'s candidate. The final column is Continue A minus Listen to B, with a participant–track bootstrap interval.

Partner comparator	Continue A (n; fraction)	Listen to B (n; fraction)	Difference [95 % CI]
PMus-Partner	174; 0.557	128; 0.547	$+0.011$ $[-0.180, 0.196]$
PMus-Partner$^+$	145; 0.510	109; 0.578	-0.068 $[-0.278, 0.160]$

Disagreements cover 33% and 31% of judgments for *PMus-Partner*, and 27% and 27% for *PMus-Partner*$^+$, under Continue A and Listen to B, respectively. Neither comparison resolves a question-specific preference. The subset was identified after data collection; a targeted study could select contrasts prospectively to distinguish continuation and response.

Context contrasts versus shared-context likelihood. We test whether the context contrast improves listener agreement over direct predictability under the shared past. The baseline ranks candidates by $-h(X \mid C)$, preferring the clip whose tokens are more predictable under the evaluator. Its orientation is fixed rather than selected from the observed choices. We compare each designated PMus measure with this baseline on the same eligible judgments. All scores are available and neither method has exact ties. Table 19 reports paired agreement differences using 10,000 shared participant–track bootstrap draws (seed 20260924). Intervals are pointwise; Holm correction covers this exploratory family of three two-sided bootstrap tail-area tests.

Table 19: Context contrasts versus conditional likelihood on identical judgments. The baseline is $-h(X \mid C)$ under the same fixed-target MusicGen evaluator. Differences are percentage points with paired participant–track 95 % intervals. All three Holm-adjusted tail areas are 0.003.

Measure / question	n	PMus	Baseline	Difference [95 % CI]
PMus-Self / Continue A	530	64.0%	44.3%	19.6 [8.6, 31.2]
PMus-Partner / Listen to B	410	63.4%	40.7%	22.7 [8.5, 37.0]
PMus-Partner$^+$ / Listen to B	410	62.2%	40.7%	21.5 [8.7, 35.1]

The context contrasts improve agreement by 20–23 percentage points over the direct conditional-likelihood ranking on the same judgments (Table 19).

Generator-only baseline. This comparison evaluates ranking against average preferences between generators. An exploratory logistic baseline uses an intercept and three generator-indicator differences, with Magenta as reference. Fitted separately for each question with one source track held out at a time, it predicts 324/530 Continue-A choices (61.1%) and 248/410 Listen-to-B choices (60.5%).

Table 20 compares PMus with this baseline on identical judgments. Pointwise participant–track intervals use 10,000 draws (seed 20260924), conditional on frozen held-out predictions and excluding training uncertainty. The paired comparisons do not resolve an accuracy difference. This is a cross-generator reference; within-generator selection is a further application to evaluate.

Adjusted choice models. *PMus-Partner* remains positively associated with listener choices after adjustment for enjoyment, generator identity, and reported instrumentation (Holm $p = 0.0224$; Table 21). We examine association and held-out prediction separately: the paired accuracy intervals leave improvements in prediction unresolved.

Separate unpenalized logistic regressions predict which clip wins each question. The unified baseline contains an intercept, the difference in Audiobox enjoyment, three generator-indicator differ-

Table 20: Exploratory practical cross-generator ranking comparison. Generator-only leave-one-source-track-out accuracy is 0.611 on Continue A ($n = 530$) and 0.605 on Listen to B ($n = 410$). Each PMus score uses its fixed orientation on the same judgments. Differences are percentage points, with pointwise 95% participant × track bootstrap intervals conditional on frozen held-out baseline predictions; they exclude model-training uncertainty and are not simultaneous.

Question	Score	PMus accuracy	PMus minus generator-only [95% CI]
Continue A	*PMus-Ctx*	0.587	-2.45 $[-14.13, +10.92]$
Continue A	*PMus-Duo*	0.662	$+5.09$ $[-4.10, +15.44]$
Continue A	*PMus-Self*	0.640	$+2.83$ $[-7.33, +14.84]$
Continue A	*PMus-Partner*	0.602	-0.94 $[-8.45, +7.84]$
Continue A	*PMus-Partner*$^+$	0.634	$+2.26$ $[-6.25, +12.37]$
Listen to B	*PMus-Ctx*	0.627	$+2.20$ $[-8.60, +13.97]$
Listen to B	*PMus-Duo*	0.668	$+6.34$ $[-4.78, +17.45]$
Listen to B	*PMus-Self*	0.663	$+5.85$ $[-5.26, +16.56]$
Listen to B	*PMus-Partner*	0.634	$+2.93$ $[-7.32, +13.57]$
Listen to B	*PMus-Partner*$^+$	0.622	$+1.71$ $[-8.72, +12.72]$

ences (Magenta is the reference), and two instrument-count covariates. The latter are the current listener's difference in multi-instrument flags and the difference in each clip's mean flag across listeners. Adding one designated PMus difference tests its incremental association on the same rows. Enjoyment and PMus differences are standardized; each PMus coefficient is the change in choice log-odds per standard deviation of the pairwise score difference, conditional on the baseline covariates. A second baseline uses enjoyment and generator identity without human instrument reports. Table 21 reports all three designated additions to the unified and no-human-report baselines.

Coefficient uncertainty uses cluster-robust sandwich covariance by participant and source track, subtracting the participant–track intersection covariance. Each component includes its finite-cluster correction. Wald tests use a normal reference and are approximate with only 22 track clusters. Tests across the three designated additions are Holm-corrected within each baseline. Using a t_{21} reference as a small-cluster sensitivity check leaves *PMus-Partner*'s unified-control association resolved (Holm $p = 0.043$), but not *PMus-Self* or *PMus-Partner*$^+$ (both $p = 0.286$). Held-out accuracy leaves each source track out, fits positive feature scales on training rows, refits both models, and pools their predictions. Baseline and full models use the same judgments. The unified baseline's instrument covariates still use human reports from held-out tracks.

Paired accuracy intervals use 10,000 shared person–track bootstrap draws (seed 20260924) on the frozen out-of-track predictions. Each draw averages the full-minus-baseline correctness difference using the same product weights. Models are not refitted within these draws, so the intervals exclude training and pipeline uncertainty.

Table 21: Item-aware adjusted choice models. Coefficients are log-odds per SD of the pairwise PMus difference, with normal-reference 95 % participant/track sandwich intervals. Holm p adjusts the corresponding coefficient tests over three measures within each baseline. Accuracies leave one source track out; Δ intervals use paired person/track resampling of frozen predictions, not model refitting. *PMus-Self* uses 522 Continue-A judgments; partner measures use 406 Listen-to-B judgments.

Added measure	Coefficient [95 % CI]	Accuracy base → full	Δ accuracy [95 % CI]	Holm p
Enjoyment + generator identity + instrument reports				
PMus-Self	0.250 $[-0.072, 0.571]$	$0.659 \to 0.669$	0.010 $[-0.026, 0.047]$	0.256
PMus-Partner	0.350 $[0.094, 0.607]$	$0.675 \to 0.700$	0.025 $[-0.022, 0.075]$	0.0224
PMus-Partner$^+$	0.306 $[-0.097, 0.710]$	$0.675 \to 0.695$	0.020 $[-0.032, 0.073]$	0.256
Enjoyment + generator identity (no human instrument reports)				
PMus-Self	0.355 $[0.022, 0.688]$	$0.628 \to 0.653$	0.025 $[-0.033, 0.084]$	0.0731
PMus-Partner	0.437 $[0.173, 0.701]$	$0.700 \to 0.672$	-0.027 $[-0.081, 0.021]$	0.0036
PMus-Partner$^+$	0.425 $[0.0004, 0.850]$	$0.700 \to 0.677$	-0.022 $[-0.071, 0.023]$	0.0731