AUDITING MUSIC EVALUATORS ACROSS ROLES: FROM CONTROL VALIDITY TO TEACHING UTILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Music evaluators assess attribute control, predict listener preferences, and provide training supervision. Evidence for one use does not by itself justify another. We develop a framework that connects evaluator selection and improvement through role-specific comparisons of control validity, direct preference quality, and teaching utility. A log-loss counterexample shows that better direct preference predictions need not yield a more beneficial one-step update to a constrained student, even with common examples and soft targets. Control evaluations show how label priors and text truncation affect measured agreement, and use waveform replacement to test whether judgments depend on the audio. For musicality preferences, complementary assessment scores improve direct prediction. Separate, fully fold-isolated teacher–student comparisons identify modest reductions in student log loss relative to matched human-only training; some gains persist after temperature calibration of both supervised and human-only students. On common human targets, comparisons among supervised students inform teacher selection, while comparisons with matched human-only training test whether supervision helps. Human-feedback adaptation lowers the log loss of MuQ-based preference scorers on held-out content relative to continued training on the original labels and temperature calibration, closing the evaluation–improvement loop for direct preference quality. The framework grounds evaluator selection and improvement in evidence matched to the intended role and, for supervision, to the student and learning procedure.

1 INTRODUCTION

A polished track may miss the requested instrument, tempo, mood, or voice. Assessing listener preference and verifying a musical condition therefore pose different evaluation problems. When an evaluator supplies training targets, its usefulness also depends on how a particular student learns from them. What evidence justifies using an evaluator in each of these roles?

Music evaluators assess generated music and can guide its improvement (Cideron et al., 2024; Ma et al., 2026; Kim et al., 2026; Zhuang et al., 2026), while multidimensional or rubric-based judgments broaden evaluation (Tjandra et al., 2025; Yao et al., 2025; Wu et al., 2026; Li et al., 2026). Evidence in one role does not establish an evaluator's validity or utility in another.

Our framework defines *control validity* by whether condition judgments are supported by audio, and *direct preference quality* by how well predicted probabilities match human preferences (Figure 1). *Teaching utility* asks whether supervision from evaluator E through interface A improves student S on evaluation set D relative to matched human-only training; it therefore depends on the learner and learning procedure.

Prior work shows that reward-model accuracy can diverge from optimization outcomes (Wen et al., 2025; Kim et al., 2025a), while distillation can depend on distribution shift (Qian et al., 2026) and teacher–learner alignment (Sucholutsky et al., 2025). We ask whether higher direct preference quality guarantees greater teaching utility for a music preference learner. A log-loss counterexample shows that it does not: with common examples and soft targets, better direct predictions can yield a less useful one-step update to a constrained student. Teaching utility requires separate evaluation.

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.

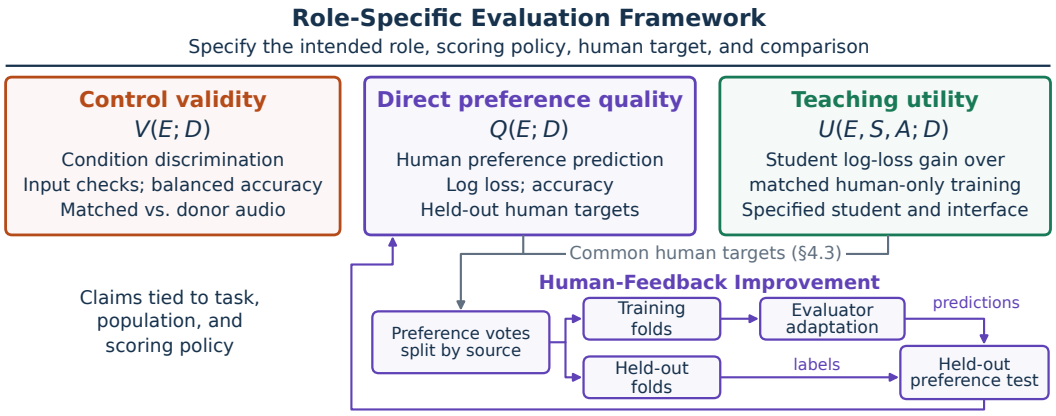

Figure 1: Role-Specific Evaluation Framework. E: evaluator; D: evaluation set; S: student; A: supervision interface. Direct-prediction and student comparisons share human targets; teaching utility is relative to matched human-only training. Source-separated feedback fitting and testing reassess direct preference quality. Claims remain specific to task, population, and scoring policy.

We align each role with its intended comparison. Control tests combine input checks, balanced scoring, and donor audio to distinguish condition discrimination from shortcuts. Preference tests assess probabilities before comparing teachers and testing gains over matched human-only training.

We adapt MuQ-based preference scorers with quality-controlled judgments and evaluate them on source-separated folds against continued training and calibration. The results reinforce the role distinctions: preprocessing and label prevalence can affect control scores, while a direct preference advantage does not establish a student advantage. Human feedback improves held-out preference prediction beyond these controls, closing the loop for a specified task, population, and scoring policy.

2 Role-Specific Evaluation Framework

The framework ties each claim to its intended use by specifying evaluator E and its scoring policy, evaluation set D, comparator, population, and unit of analysis. Control validity and direct preference quality assess the evaluator's judgments; teaching utility assesses their contribution to student S through supervision interface A. Feedback closes the loop for direct preference quality by fitting on training judgments and reassessing on held-out judgments.

Control validity: separating condition judgments from shortcuts. For audio x_i, let c_i^+ be the human-supported condition and c_i^- its alternative. A scorer $s_E(x, c)$ induces margin

$$m_i^E(x) = s_E(x, c_i^+) - s_E(x, c_i^-). \tag{1}$$

Correctness is $\mathbf{1}\{m_i^E(x_i) > 0\}$; zero margins receive no credit. The condition contrast must first survive preprocessing: identical effective inputs cannot represent the intended distinction. Balanced accuracy (BA) averages recall for the two human label classes: source or edited condition. To test waveform dependence, we preserve the conditions and human orientation but replace x_i with a deterministic donor x_i^{don} from another prompt, recording, and attribute family. The matched-minus-donor raw accuracy gap is

$$G_D(E) = \frac{1}{|D|} \sum_i \left[\mathbf{1}\{m_i^E(x_i) > 0\} - \mathbf{1}\{m_i^E(x_i^{\text{don}}) > 0\} \right]. \tag{2}$$

After this input check, we use $V(E; D) = (\text{BA}_D(E), G_D(E))$ as a validity profile. Let $0 < \pi < 1$ be the fraction of items where humans favor the source condition. An audio-blind majority rule attains raw agreement $\max\{\pi, 1 - \pi\}$, BA $1/2$, and $G_D = 0$. Balanced accuracy and donor replacement thus diagnose different failures. Under the fixed policy, a positive sample gap indicates waveform dependence; population claims require uncertainty assessment. Donors alter multiple acoustic factors, so a gap does not establish causal recognition of the named attribute.

Direct preference quality. For an ordered, possibly conditioned audio pair z, $y \in [0, 1]$ is the human target for its first candidate and $p_E(z) \in (0, 1)$ is the corresponding evaluator probability. Negative log-likelihood (NLL) is fractional binary log loss:

$$L_D(E) = \frac{1}{|D|} \sum_{(z,y)\in D} \left[-y \log p_E(z) - (1 - y) \log(1 - p_E(z))\right]. \tag{3}$$

Logarithms are natural. We define $Q_{\mathrm{NLL}}(E; D) = -L_D(E)$, so higher is better; log loss is strictly proper (Gneiting & Raftery, 2007). Accuracy, Q_{acc}, uses prespecified hard-label and tie rules and is a separate endpoint: identical hard decisions can have different probability losses. Calibration parameters must be fit without evaluation labels. Calibration diagnostics complement rather than replace evaluation under the proper score (Guo et al., 2017).

Teaching utility: measuring a change in the student. Interface A fixes shared training and prediction rules, including aggregation and training-side calibration, and maps evaluator outputs to examples, targets, and weights. For student specification S, let S_A^E denote the predictor trained with E's pseudo-supervision and S_A^0 its schedule-matched human-only counterpart with no pseudo-loss (ZERO). We define

$$U(E, S, A; D) = L_D(S_A^0) - L_D(S_A^E). \tag{4}$$

For a held-out evaluation set D, positive utility is a reduction in student NLL. Initialization, human supervision, optimization, and ensemble size are held fixed. Utility measures the specified pseudo-supervision procedure within this configuration, not each teacher's best performance under separate tuning.

Why the supervision interface matters. For human risk R_H and weighted pseudo-risk $R_{E,A}$, one gradient step $\theta'_E = \theta - \eta \nabla R_{E,A}(\theta)$ gives, for locally twice continuously differentiable R_H as $\eta \to 0$,

$$R_H(\theta) - R_H(\theta'_E) = \eta \langle \nabla R_H(\theta), \nabla R_{E,A}(\theta) \rangle + O(\eta^2). \tag{5}$$

Selection, target conversion, weights, and student parameterization affect this gradient alignment, so direct log loss alone gives no monotonic utility guarantee. The following construction isolates the distinction without selection, weighting, or distribution shift.

Proposition 1 (Higher direct preference quality need not imply greater one-step utility)**.** *Even on the same population, with common examples, unit weights, a shared soft-target interface, and a single gradient step with the same step size, there exist evaluators E_1, E_2 and a student S such that $Q_{\mathrm{NLL}}(E_1; D) > Q_{\mathrm{NLL}}(E_2; D)$ but $U(E_1, S, A; D) < 0 < U(E_2, S, A; D)$.*

The constant-probability student responds only to the mean target. An evaluator with better direct predictions can still place that mean on the wrong side of the current student prediction, inducing an update away from the human optimum. Appendix E.1 gives the construction and proof.

This existence result motivates separate utility measurement but does not predict the hard-target results below. Further derivations are in Appendix E.

Applying the framework. Control validity uses human-supported conditions; preference quality and utility use human preference targets. Held-out claims require excluding evaluation labels from fitting, calibration, and schedule selection, including during feedback adaptation (Appendix D.6). Different tasks and scoring policies preclude a joint evaluator ranking across roles.

3 Evaluating Fine-Grained Control

Control validity tests whether an evaluator selects the condition that better describes a recording. We pair single-attribute comparisons with checks for label-prior shortcuts and loss of the condition contrast at the model input.

This design follows compositional audio–language evaluation (Ghosh et al., 2024; Chen et al., 2026; Wang et al., 2026) and work separating conditioning adherence from general quality (Heydari, 2026).

Each item pairs a recording with source and counterfactual conditions differing in one attribute; "source" denotes provenance, not correctness. Three distinct listeners select the source condition, the

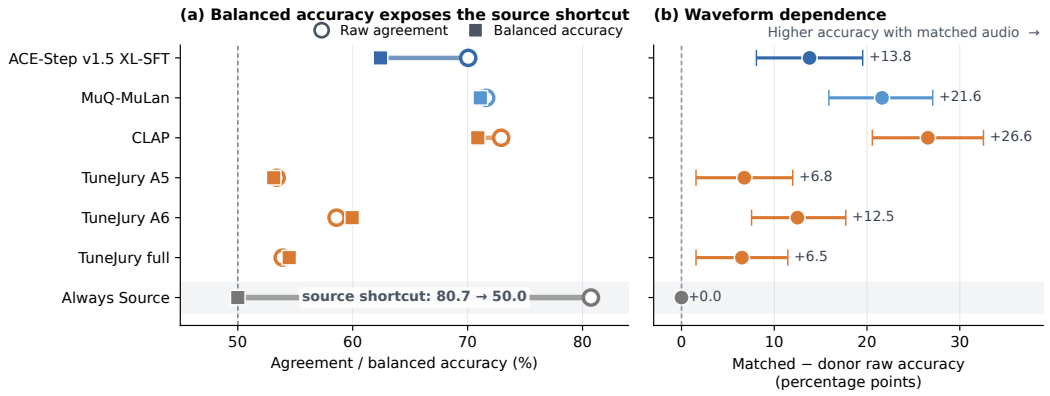

Figure 2: Control validity on the 384 initial human-accepted condition pairs. (a) Raw agreement (circles) and balanced accuracy (BA; squares); Always Source ignores audio. (b) G_D is the matched-minus-donor raw-accuracy gap after replacing each recording with a donor whose prompt, recording, and attribute family all differ from the target's, while holding conditions fixed; bars are nominal 95% prompt-bootstrap confidence intervals. Blue marks the prespecified ACE-Step evaluation and MuQ-MuLan control, orange the exploratory policies, and gray the audio-blind diagnostic. TuneJury A5 uses CLAP audio and text, A6 uses MERT audio with CLAP text, and full combines both audio representations with CLAP text.

counterfactual, both, neither, or unsure. A pair is retained when at least two listeners select the same single condition, yielding labels for 384 of 480 CMI-Pref-Pseudo pairs across six attribute families. Orientation-reversed repeats check consistency but do not contribute labels (Appendix B).

We compare an ACE-Step v1.5 XL-SFT probe trained on rule-derived targets (Gong et al., 2026), zero-shot MuQ-MuLan (Zhu et al., 2025) and CLAP (contrastive language–audio pretraining) (Wu et al., 2023), and TuneJury preference heads (Kim et al., 2026) using CLAP/MERT (Li et al., 2024) features. Probe training excludes evaluation prompts and recordings; no policy uses evaluation labels for fitting or calibration. CLAP uses one deterministic audio crop for both texts. Comparisons concern complete scoring policies, whose training data and temporal coverage differ.

3.1 Separating label prevalence from waveform dependence

Accepted labels are imbalanced despite balanced edit directions: 310 source and 74 counterfactual. Always Source ignores audio and selects the unedited condition, achieving the highest raw agreement (80.73%) but BA of 50% and $G_D = 0$. MuQ-MuLan and CLAP lead the BA point estimates (Figure 2); raw agreement alone does not establish condition sensitivity.

Donors follow Section 2, are selected without labels or model scores, and match generator identity where possible. For ACE-Step's prespecified test, the lower bounds of the aggregate prompt-bootstrap 95% confidence intervals (CIs) exceed 50% for matched-audio raw accuracy and zero for G_D. At the attribute-family level, only mood and vocal gender meet both criteria, short of the prespecified requirement of at least four of the six families. The aggregate result therefore does not establish the prespecified breadth of condition discrimination across attribute families. For CLAP, generator-cluster intervals for G_D remain positive on the initial and same-source follow-up sets after matching coarse amplitude and spectral statistics. These exploratory checks support waveform dependence but do not isolate the edited attribute's contribution (Appendices A and C).

3.2 Condition visibility and external transfer

A sample-disjoint CMI-Pref-Pseudo follow-up tests stability of the CLAP–TuneJury comparison; Music Arena (Kim et al., 2025b) tests transfer to another data source, excluding exact prompt overlap with CMI-Pref-Pseudo. Policies were frozen before returned labels or outputs were observed, and the primary TuneJury head was trained without Music Arena. Upstream encoder-pretraining overlap cannot be fully audited. The same-source collection required a model-blind coverage extension

before inference; without a predeclared sequential adjustment, it provides sensitivity rather than confirmatory evidence.

The prespecified external evaluation tests three distinct claims: above-chance CLAP discrimination, waveform dependence, and an advantage over the primary TuneJury comparator. Joint success requires the lower bounds of the 95% CIs to exceed 50% for CLAP BA and zero for both G_D and the CLAP–TuneJury BA difference. Paired bootstrap intervals cluster by prompt for the initial and same-source sets and by user externally. External BA resamples users within label classes; a post-hoc whole-user sensitivity preserves dependence across classes. Secondary same-source attribute tests use Holm correction; displayed CIs remain nominal (Appendix A).

CLAP's BA point estimate exceeds that of full TuneJury on the initial and same-source sets and that of the head trained without Music Arena on the external set (Table 1). All three components of the joint criterion pass under the prespecified external resampling (Table 1). In the whole-user sensitivity, the donor-gap and CLAP–TuneJury BA-difference intervals remain above zero, but the CLAP BA interval, $[49.93, 64.81]\%$, includes 50% (Appendix A.7). Thus the relative policy and waveform contrasts persist under both schemes, but the above-chance component and joint claim depend on how users are resampled. The comparison tests zero-shot CLAP against a preference-trained head transferred to condition discrimination.

Table 1: CLAP control discrimination across three human-accepted sets. BA is balanced accuracy (%); G_D is matched-minus-donor accuracy with conditions fixed. Differences are percentage points; brackets are nominal 95% bootstrap confidence intervals, clustered by prompt for the initial and same-source sets and by source user for external Music Arena (MA). Initial and same-source comparisons use full TuneJury and provide exploratory and sensitivity evidence, respectively; external MA uses the head trained without MA. Unbracketed initial entries are descriptive point estimates. External BA resamples users within label classes; its post-hoc whole-user interval is $[49.93, 64.81]\%$ and includes 50%.

Evaluation	Accepted items	CLAP BA	G_D	CLAP–TuneJury BA diff.
Initial	384	70.88	+26.56 [20.57,32.55]	+16.40
Same-source	219	64.93 [56.46,73.14]	+21.46 [13.70,28.77]	+15.48 [4.50,26.47]
External MA	359	57.30 [50.51,64.36]	+21.73 [15.57,28.15]	+14.09 [6.30,21.91]

Text truncation also limits interpretation: CLAP's 77-token limit makes condition inputs identical for 82 of 359 items, accounting for every matched- and donor-audio tie; BA gives these ties no credit. CLAP retains a descriptive BA advantage on the 277 distinguishable pairs. This post-hoc input diagnosis supplements the full-set evaluation (Appendix A.7).

Agreement-stratified comparisons and varying acceptance rates further delimit the accepted-label population; they do not establish that ambiguity explains performance differences (Appendix Tables 16 and 5). Control validity therefore depends on both the scoring policy and the accepted-label population. Listener preference quality requires separate evaluation on human preference targets.

4 Direct Preference Quality and Teaching Utility

For preference prediction, the framework distinguishes the quality of an evaluator's own probabilities from the benefit of its supervision to a student. We first ask whether combining assessment scores improves prediction of human musicality preferences, then whether supervision from these scorers improves a student relative to matched training on human labels alone. Evaluating competing scorers and their matched students on the same human targets then tests whether the better direct predictor also provides more useful supervision.

4.1 Complementary assessments improve direct prediction

We test whether assessments of multiple musical aspects provide additional information for predicting human musicality preferences. External 4D combines AudioBox's track-average Content Enjoyment and Production Quality scores with the TuneJury reward and MuQ-Eval's Musical Impression score (Zhu & Li, 2026); D counts score inputs. To capture additional aspects and variation within a track, the

10D fusion uses four AudioBox aspect scores, each summarized by its mean and standard deviation across audio windows, together with the TuneJury and MuQ-Eval scores. Adding SongEval's five and SongBench's seven aspect scores yields 22D (Appendix D.1). For each fusion, we scale candidate-score differences by their root mean squares over fitting data and learn their weights from human preferences with a no-intercept logistic model. For comparison, the AudioBox baseline averages its two scaled score differences and fits a single logistic coefficient.

Development evaluation uses 3,527 individual CMI-RewardBench Musicality votes (Ma et al., 2026), grouped by shared prompt or audio identity into five folds. Each fold's predictions use scales and fusion weights fitted outside that fold. Selection among the SongEval and SongBench input combinations is nested within each outer training set; all five folds select the complete 22D combination. For these preference comparisons, TuneJury uses a head trained without Music Arena and a fixed zero text vector, so its score depends on audio alone.

Fusion models and the AudioBox calibrator are fitted on all development votes, then frozen for Music Arena and the official CMI-RewardBench test. Music Arena's overall preferences test transfer across source and target; the official test retains Musicality. Neither supplies fitting or selection labels; the official test is evaluated once (Table 2).

Expanding from 10D to 22D lowers development NLL and yields distinct improvements on the two independent evaluations (Table 2). On Music Arena, 22D raises accuracy relative to 10D by 2.16 percentage points (95% user-cluster CI $[0.90, 3.39]$). On the official Musicality test, NLL falls by 0.0321 (95% component-bootstrap CI $[0.0120, 0.0518]$), a 6.15% reduction; the interval for the accuracy difference includes zero. The independent comparisons therefore support higher accuracy on Music Arena and lower NLL on the official CMI-RewardBench Musicality test.

Development ablations give lower NLL point estimates for full SongEval/SongBench vectors than for their Musicality scores alone; learned fusion also exceeds the tested averaging and voting rules in accuracy (Appendix Tables 25 and 26). These comparisons support combining correlated assessment scores for direct prediction.

Table 2: Direct preference prediction on CMI-RewardBench and Music Arena. CMI development (3,527 votes) and official test (500 pairs) measure Musicality preferences; Music Arena (MA; 3,660 decisive comparisons) measures overall user preference. Development predictions are out of fold; MA and official predictions use frozen fits on all development votes. D counts fused score inputs. Accuracy is in percent; negative log likelihood (NLL) is lower-is-better. Bold marks column leaders, not significance. Scoring details are in Appendix D.1.

	Development		MA	Official	
Evaluator	**Acc.**	**NLL**	**Acc.**	**Acc.**	**NLL**
AudioBox	73.46	0.52891	59.81	73.80	0.53320
External 4D	75.39	0.50417	63.63	73.80	0.52682
10D fusion	75.90	0.49881	66.01	73.80	0.52191
22D fusion	**77.43**	**0.48201**	**68.17**	**75.40**	**0.48982**

4.2 Measuring supervision gains against matched training

We test supervision against matched human-only training using an existing pair pool's training split, audio-disjoint from validation, discarding its legacy labels and confidence scores. External 4D selects confident pairs where its three evaluators agree with the fused preference; other teachers match its pair count using their own calibrated confidence. Probabilities supply hard targets $\hat{y}_i = \mathbf{1}\{p_E(z_i) \geq 0.5\}$ and weights $w_i = 2|p_E(z_i) - 0.5|$. Utility therefore compares complete procedures—selection, targets, and weights—rather than predictions alone (Appendix D.8).

Each teacher supervises the same MuQ student architecture (Q0), which uses frozen representations from twelve encoder layers and trains a fusion module, adapters, and task heads. Joint training adds a Musicality pseudo-label loss to the human task losses, followed by human-only fine-tuning. ZERO matches initialization, seeds, and training steps with zero pseudo-loss contribution (Appendix D.1).

Fold isolation excludes evaluation labels from teacher fitting as well as student training. In each outer fold, inner-training teacher fits generate pseudo supervision; an External-4D student's inner-validation

performance selects one schedule for all students and ZERO. Teachers and pseudo supervision are then rebuilt on outer-training data, and students are trained afresh. Pair counts match within each fitting scope; out-of-fold (OOF) probabilities average five seeds that exclude the evaluation fold. This post-hoc comparison fixes teacher features and student architecture. Earlier development results used teachers fitted on all development labels and remain conditional diagnostics (Table 21; Appendix D.4).

All four supervised ensembles have 0.62–0.97% lower development NLL than ZERO, with little accuracy change; four or five of five seeds improve for each teacher (Table 3). Nominal, unadjusted intervals condition on fitted ensembles and exclude full-retraining uncertainty.

Table 3: Fully fold-isolated MuQ student effects on 3,527 development votes before student temperature calibration. NLL is negative log likelihood; changes subtract matched human-only ZERO (NLL 0.48765). Negative ΔNLL favors supervision; accuracy changes are percentage points. Nominal 95% intervals use component resampling within each outer fold and condition on fitted five-seed ensembles. D counts teacher inputs; better seeds have lower NLL than matched ZERO.

Teacher	**ΔNLL [95% CI]**	**ΔAcc.**	**Better seeds**
AudioBox	-0.00472 [-0.00812, -0.00128]	$+0.14$	5/5
External 4D	-0.00327 [-0.00634, -0.00025]	-0.03	4/5
10D	-0.00474 [-0.00808, -0.00140]	$+0.09$	4/5
22D	-0.00302 [-0.00569, -0.00042]	$+0.00$	4/5

To test gains beyond scalar confidence adjustment, we calibrate each student and ZERO with a separately fitted positive temperature, preserving threshold decisions. In each fold, temperatures fitted to inner-ensemble validation predictions transfer to the corresponding outer ensembles. AudioBox and 10D students retain NLL gains of 0.00373 and 0.00409 over calibrated ZERO ($p_{\mathrm{Holm}} = 0.0384$ and 0.0228 across four comparisons); External 4D and 22D do not pass that correction. Some gains thus persist under symmetric calibration, conditional on fitted ensembles and transferred temperatures (Appendix D.5).

Separate evaluation uses the original frozen 25-member student ensembles, distinct from these refits, on three-listener judgments unused in teacher or student fitting. External-4D supervision lowers NLL relative to ZERO and self-training using the MuQ student's pseudo-labels; paired 95% intervals exclude zero. Against ZERO, NLL falls by 0.64% on valid-vote means, while the majority-accuracy point estimate falls by 0.71 points on pairs with three valid votes (Appendix D.11). The predeclared student–ZERO NLL comparisons on the official benchmark remain inconclusive (Appendix Figure 4).

These results identify settings in which external supervision improves probability prediction over matched human-only training.

4.3 Comparing roles on common human targets

To compare direct preference quality with teaching utility, we evaluate four scorers and their matched MuQ students on the same audio pairs and human targets. Because all students share the same human-only control (ZERO), lower student NLL implies greater teaching utility (Equation 4); comparisons with ZERO separately test whether supervision helps.

Each of 2,400 pairs is assigned to four listeners from a pool of 24. Pair identities differ from the earlier listening evaluation, but source components overlap. Scorer predictions and the original 25-member student ensembles are frozen before these judgments; the fold-isolated refits form a separate comparison.

Quality control (QC) uses annotation records without model predictions. Retained pair targets average accepted votes encoded as 1 (A), 0 (B), and 0.5 (tie), omitting cannot-judge responses; NLL weights pairs equally. The original analysis retains 999 pairs. Two post-hoc reannotation rounds replace designated annotators' earlier returns; the unchanged QC rules expand coverage to 2,397 pairs. These versions provide sensitivity evidence within the same annotation pool (Appendices D.12–D.13).

The prespecified primary comparison tests a reversal between the direct and student orderings of External 4D and AudioBox. It requires a negative External-4D-minus-AudioBox direct NLL difference and a positive matched-student difference, with both 95% intervals excluding zero. The

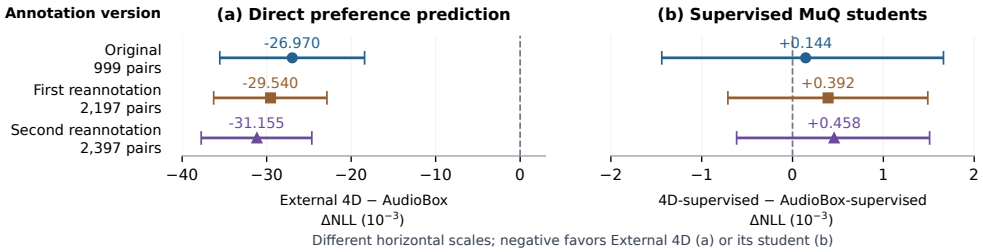

Figure 3: Direct-prediction and student negative log likelihood (NLL) differences on common human targets. Panel (a) shows External 4D minus AudioBox NLL; panel (b) shows the same contrast between their supervised MuQ students. Negative differences favor External 4D (a) or its student (b). External 4D fuses four score outputs. Bars are nominal 95% source-component bootstrap confidence intervals. The annotation versions contain overlapping pairs and use unchanged model predictions; the panels have different axis ranges.

95% interval for the direct NLL difference lies entirely below zero, whereas the matched-student interval spans zero (Figure 3). Both post-hoc reannotation versions preserve this pattern, including in analyses restricted to the pairs shared across all three versions. These results support External 4D's direct prediction advantage but leave its teaching utility relative to AudioBox unresolved. They establish neither the prespecified ranking reversal nor equivalent teaching utility for these two teachers.

The secondary comparisons show lower NLL for the 10D-supervised MuQ student than for the AudioBox- and 22D-supervised students in all three annotation versions. Holm correction is applied separately within each version to the seven student comparisons specified for the original analysis. The 10D student's gain over the matched human-only control (ZERO) is supported in both post-hoc reannotation versions, but remains inconclusive in the original analysis (Appendix D.13). We next test whether the accepted judgments from the final version can improve the 10D-supervised and human-only students as preference scorers, separating feedback training from evaluation by source component.

5 Improving Evaluators with Human Feedback

We test whether human feedback improves the MuQ students' direct preference quality on held-out judgments beyond continued training on the original labels or scalar calibration. We use the final annotation version, comprising 9,020 valid overall-musicality votes on 2,397 pairs from 23 QC-accepted annotators.

Learning from human preferences. To test adaptation with and without prior teacher supervision, we start from the original 10D-supervised and human-only (ZERO) MuQ ensembles evaluated on common human targets. For Bradley–Terry training, each new pair's soft target is the mean of its valid overall-musicality votes, with ties encoded as 0.5 and unjudgeable responses omitted. Each ensemble member minimizes $(1-\alpha)L_{\text{old}} + \alpha L_{\text{new}}$, where α weights feedback, L_{old} is log loss on replayed human Musicality votes, and L_{new} averages log loss equally over new training pairs using their soft targets. Because log loss is affine in its target, each new pair's loss equals the average of its valid-vote losses, with equal weight across new pairs. Layer fusion, the audio adapter, and the Musicality head are updated; the MuQ-MuLan audio encoder and projection remain frozen.

Testing improvement on held-out content. We first test prediction on new content within the same annotator pool. The feedback data share no source components with pseudo-training data and no audio hashes with the original human development data. We divide the feedback pairs into five new outer folds by source component. For each held-out fold and initialization, all 25 members are adapted using feedback only from the other four folds, and their probabilities are averaged for evaluation.

Comparisons with frozen predictions measure change from the starting models; old-data continuation controls for further training on the original labels, while temperature calibration tests global confidence adjustment. Inner validation within the training folds selects the feedback schedule, which continuation follows with the same learning rate and number of optimizer steps. For each initialization and outer fold, a positive temperature rescales the logit of the frozen ensemble probability and is fitted by minimizing NLL on the four training folds. Because the adaptation design was chosen after inspecting the common-target judgments and model comparisons, these estimates are exploratory; intervals condition on fitted predictions (Appendix D.14).

Table 4: Five-fold content evaluation of MuQ feedback adaptation on 2,397 pair-mean targets. NLL is negative log likelihood (lower is better). Starts are human-only (ZERO) or 10D-supervised; D counts teacher inputs. Predictions average 25 members' probabilities. Old-data training replays original labels; temperature-only fits a scalar; feedback adds new training-fold judgments to replay. Fits exclude test-source components. Bold marks row minima; conditional intervals are in Appendix D.14.

Initialization	Frozen	Old data only	Temperature only	Feedback + replay
ZERO start	0.4138	0.4341	0.4073	**0.3838**
10D start	0.4104	0.4277	0.4060	**0.3813**

For both initializations, feedback lowers held-out NLL by about 7% relative to the corresponding frozen ensemble (Table 4). Feedback outperforms all three controls at both initializations in NLL, with one Holm correction over the six comparisons ($p_{\mathrm{Holm}} < 0.001$). These results support improved preference probability prediction on held-out content within this annotator pool.

Holding out listeners and content. We next ask whether the improvement extends to judgments from listeners excluded from feedback fitting. We restart both ensembles from their pre-feedback models, using feedback targets recomputed solely from 17 annotators' judgments and excluding the other six from feedback fitting, selection, and calibration. Testing uses one fixed partition of 473 pairs on held-out source components, each with one valid judgment from a reserved listener (Appendix D.15). On this fixed partition, mean NLL is lower than that of the frozen scorers for both initializations.

Human-feedback MuQ reward model. The audio-only MuQ reward model packages the human-only (ZERO) ensemble adapted in the listener-and-content holdout, including the shared frozen encoder and projection with all 25 adapted members. It averages member probabilities; its mean single-audio score is an uncalibrated relative utility (Appendix D.15).

6 Conclusion

Music evaluator quality depends on the intended use. Our framework distinguishes control validity, direct preference quality, and teaching utility for a specified student and learning procedure. The log-loss counterexample shows that better direct preference prediction need not yield a more beneficial one-step update to a constrained student, even on common examples with soft targets.

On the control task, an audio-blind rule can attain high raw agreement under imbalanced labels, and text truncation can erase the condition contrast at the model input. Donor comparisons complement input and label-prior checks by testing waveform dependence. For preferences, combining complementary assessments improves direct prediction, while separate, fully fold-isolated comparisons identify reductions in student NLL relative to matched human-only training. Within the evaluated annotator pool, adapting MuQ-based preference scorers with human judgments yields lower NLL on held-out content than continued training on the original labels or temperature calibration of the corresponding frozen ensembles. The framework thus connects role-specific evaluator selection with human-feedback improvement, grounding each claimed benefit in evidence for its intended role. A next step is to compare the teaching utility of the pre-feedback and feedback-adapted MuQ scorers using matched students and a common human-only control.

Ethics, Data Governance, and Human Evaluations

Participant Recruitment and Labor Standards. A specialist data-annotation vendor organized annotation, review, and quality control for judgments of generated music. Recruitment records for the initial control evaluation and the independent student test (T7) cover 12 contributors with bachelor's or master's degrees in music performance, musicology, composition, or music education and 1–10 years of relevant experience. The vendor confirmed hourly compensation substantially above the applicable statutory minimum for comparable non-full-time work and compliance with local labor requirements. Aggregate service records are retained for review, with identifying vendor information redacted during anonymous assessment.

Participant Tracking and Unit of Analysis. The initial control annotation involved all 12 contributors; T7 used nine from that pool, so these participant counts are non-additive. The same-pair study (T29) and both reannotation rounds (R2 and R3) drew on the same pool of 24 distinct contributors, with one person assigned to each worker slot. Analyses use item- or pair-level outcomes with task-specific clustering. The appendices distinguish participant counts, task assignments, repeated judgments, and quality-control criteria.

Participant Information and Privacy. The research team explained the listening study's purpose to the provider and asked it to convey that purpose when recruiting annotators. The provider's ethical review of the study preceded data collection. The provider recruited participants and described the listening and annotation tasks they would perform. Reported analyses use anonymous task and worker identifiers and omit personal names and free-text notes. Administrative provenance is retained under restricted access.

Evaluation Blinding and Protocol Validity. Task interfaces concealed generator identities, private source/orientation keys, legacy labels, and model predictions as applicable, while displaying the alternatives needed for each judgment. For the first reannotation, the project owner confirms relistening of assigned repair packets, particularly those with playback and consistency failures. Both reannotation rounds are analyzed as post-hoc rework on the same tasks. The official benchmark was evaluated once under a frozen protocol; its results did not change fitting, calibration, filtering, or candidate selection. Music judgments depend on cultural and listening context, and the measured scores do not replace human judgment in consequential creative decisions.

Asset Licensing and Audio Redistribution Governance. Archived model cards and manifests declare CMI-Pref and CMI-Pref-Pseudo as CC BY-NC-SA 4.0; Music Arena as CC BY 4.0; CMI-RM, MuQ-MuLan weights, MERT-v1-330M, and TuneJury heads as CC BY-NC 4.0; LAION-CLAP-Music as CC0 1.0; SongEval as CC BY-NC-SA 4.0; SongBench as research-only noncommercial; ACE-Step code/checkpoint as MIT; and MuQ code as MIT. Raw-audio redistribution is subject to item-level rights review in addition to these asset declarations. The reproduction packages contain aggregate summaries, de-identified preference targets, fixed evaluation probabilities, and component/fold keys. They exclude audio, prompts, worker identities, and raw annotation returns. The packages are prepared locally. To preserve author anonymity, public release is deferred until the anonymous review process has concluded. Any release of code, model weights, or derived research artifacts additionally requires internal company approval and will be limited to the approved scope and applicable licenses and permissions.

Reproducibility statement

The appendices document control-task construction and scoring (Appendices A and C), preference data use and fold-scoped fitting (Appendices D.3 and D.4), original T29 and both reannotation analyses (Appendices D.12–D.13), human-feedback adaptation (Appendices D.14 and D.15), and the theory's assumptions and derivations (Appendix E). Appendix D.6 maps the reproducible calculations to seven locally prepared packages in `ICLR2027_REPRODUCTION_SUPPLEMENT_20260922.zip`: an aggregate worked example; original-T29 pair-level statistics; vote-level statistics for the fully fold-isolated study (T30) with recorded fitting-scope checks; training-side temperature fitting and statistical replay for the calibration follow-up; the original/first/second reannotation statistical comparison; fixed-OOF statistics for human-feedback adaptation; and the separate listener-and-content holdout statistics. Commands, dependencies, and hashes are included. These reconstruct calculations and check recorded memberships; they do not replay annotation, feature extraction, or teacher/student training. The full experiment archive is retained locally and is not included in this

limited supplement. The complete human-feedback MuQ reward model described in Section 5 and its standalone inference code are prepared as a separate local package. The manuscript includes methods, results, and calculation instructions; access to these local packages for anonymous reviewers has not been established. Public release follows the timing, company-approval, scope, and rights conditions stated above.

AI USE STATEMENT

The human authors set the core research question, defined the role distinction (V, Q, U), and retained final authority over protocol design, hypotheses, and evidential interpretation.

OpenAI Codex was used as a research assistant to implement and debug experiment, audit, statistical, and figure code, assist with hypothesis and protocol refinement, analysis, interpretation, checking mathematical arguments, and drafting, and execute the predeclared official-test inference under explicit authorization after the protocol hash was frozen. It did not generate human preference annotations, and it did not fit, calibrate, select, or suppress any model using official-test outcomes.

Checks of AI-assisted work include tests, hash verification, and rerunnable analyses. The project owner authorized official-test use; the archived protocols and amendments record gate definitions and their timing.

Human authors are responsible for independently verifying citations, claims, human-subject facts, licenses, code, and artifacts before submission. The AI system is not an author.

REFERENCES

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.2307/2334029.

Chuyang Chen, Bea Steers, Brian McFee, and Juan Bello. Evaluating compositional structure in audio representations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026. URL https://arxiv.org/abs/2603.13685.

Geoffrey Cideron, Sertan Girgin, Mauro Verzetti, Damien Vincent, Matej Kastelic, Zalán Borsos, Brian McWilliams, Victor Ungureanu, Olivier Bachem, Olivier Pietquin, Matthieu Geist, Léonard Hussenot, Neil Zeghidour, and Andrea Agostinelli. MusicRL: Aligning music generation to human preferences. *arXiv preprint arXiv:2402.04229*, 2024. URL https://arxiv.org/abs/2402.04229.

Chengqian Gao, Haonan Li, Liu Liu, Zeke Xie, Peilin Zhao, and Zhiqiang Xu. Principled data selection for alignment: The hidden risks of difficult examples. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18386–18409, 2025. URL https://proceedings.mlr.press/v267/gao25f.html.

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Evuru, S. Ramaneswaran, S. Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. CompA: Addressing the gap in compositional reasoning in audio-language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/43c18853329c7504996b255252b6cb1f-Abstract-Conference.html.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Junmin Gong, Yulin Song, Wenxiao Zhao, Sen Wang, Shengyuan Xu, Jing Guo, and Xuerui Yang. ACE-Step 1.5: Pushing the boundaries of open-source music generation. *arXiv preprint arXiv:2602.00744*, 2026. URL https://arxiv.org/abs/2602.00744. Preprint; not peer reviewed.

Florian Grötschla, Ahmet Solak, Luca A. Lanzendörfer, and Roger Wattenhofer. Benchmarking music generation models and metrics via human preference studies. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10887745. URL https://ieeexplore.ieee.org/document/10887745/.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Morteza Heydari. How well do generative music models follow emotion conditioning? *arXiv preprint arXiv:2608.29987*, 2026. URL https://arxiv.org/abs/2608.29987. Preprint; not peer reviewed.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Sunghwan Kim, Dongjin Kang, Taeyoon Kwon, Hyungjoo Chae, Dongha Lee, and Jinyoung Yeo. Rethinking reward model evaluation through the lens of reward overoptimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13252–13280, 2025a. doi: 10.18653/v1/2025.acl-long.649. URL https://aclanthology.org/2025.acl-long.649/.

Yonghyun Kim, Wayne Chi, Anastasios N. Angelopoulos, Wei-Lin Chiang, Koichi Saito, Shinji Watanabe, Yuki Mitsufuji, and Chris Donahue. Music arena: Live evaluation for text-to-music. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/4962124728954e41e3cfe990b3a307f4-Abstract-Creative_AI_Track.html. Creative AI Track.

Yonghyun Kim, Junwon Lee, Haiwen Xia, Yinghao Ma, Junghyun Koo, Koichi Saito, Yuki Mitsufuji, and Chris Donahue. TuneJury: An open metric for improving music generation preference alignment. *arXiv preprint arXiv:2606.17006*, 2026. URL https://arxiv.org/abs/2606.17006.

Klaus Krippendorff. Computing Krippendorff's alpha-reliability. Technical report, Annenberg School for Communication, University of Pennsylvania, 2011. URL https://www.asc.upenn.edu/sites/default/files/2021-03/Computing%20Krippendorff%27s%20Alpha-Reliability.pdf. Literature updated September 13, 2013.

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2206–2217, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Li_Your_Diffusion_Model_is_Secretly_a_Zero-Shot_Classifier_ICCV_2023_paper.html.

Haitao Li, Tian Tan, Yuguang Yang, Shan Yang, and Xie Chen. AnyAudio-Judge: A dynamic rubric-based benchmark and evaluator for audio instruction following. *arXiv preprint arXiv:2606.03116*, 2026. URL https://arxiv.org/abs/2606.03116.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. MERT: Acoustic music understanding model with large-scale self-supervised training. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/33dffa2e3d2ab74a783d1a8c292f66d9-Abstract-Conference.html.

Cheng Liu, Hui Wang, Jinghua Zhao, Shiwan Zhao, Hui Bu, Xin Xu, Jiaming Zhou, Haoqin Sun, and Yong Qin. MusicEval: A generative music dataset with expert ratings for automatic text-to-music evaluation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10890307. URL https://ieeexplore.ieee.org/document/10890307/.

Yinghao Ma, Haiwen Xia, Hewei Gao, Weixiong Chen, Yuxin Ye, Yuchen Yang, Sungkyun Chang, Mingshuo Ding, Yizhi Li, Ruibin Yuan, Simon Dixon, and Emmanouil Benetos. CMI-RewardBench: Evaluating music reward models with compositional multimodal instruction. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2603.00610.

Haolong Qian, Xianliang Yang, Yinuo Ma, Lirong Che, Feng Lu, Ye Guo, Lei Song, Jiang Bian, and Chun Yuan. The quality-utility paradox: Why high-reward data impairs small model mathematical reasoning. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.16152.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *The VLDB Journal*, 29:709–730, 2020. doi: 10.1007/s00778-019-00552-1.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/06964dce9addb1c5cb5d6e3d9838f733-Abstract.html.

Ilia Sucholutsky, Katherine M. Collins, Maya Malaviya, Nori Jacoby, Weiyang Liu, Theodore R. Sumers, Michalis Korakakis, Umang Bhatt, Mark Ho, Joshua B. Tenenbaum, Zachary A. Pardos, Adrian Weller, and Thomas L. Griffiths. Representational alignment supports effective teaching. In *Proceedings of the Innovation and Responsibility in AI-Supported Education Workshop*, volume 273 of *Proceedings of Machine Learning Research*, pp. 146–173, 2025. URL https://proceedings.mlr.press/v273/sucholutsky25a.html.

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.

Jinting Wang, Yuguang Yang, Shengyu Li, Yan Rong, Shan Yang, Xiaoda Yang, and Li Liu. AudioScape-TTA: A structured soundscape benchmark for fine-grained text-to-audio evaluation. *arXiv preprint arXiv:2608.04479*, 2026. URL https://arxiv.org/abs/2608.04479. Preprint; not peer reviewed.

Xueru Wen, Jie Lou, Yaojie Lu, Hongyu Lin, Xing Yu, Xinyu Lu, Ben He, Xianpei Han, Debing Zhang, and Le Sun. Rethinking reward model evaluation: Are we barking up the wrong tree? In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/93f250215e4889119807b6fac3a57aec-Abstract-Conference.html.

Dapeng Wu, Shun Lei, Wei Tan, Guangzheng Li, Yunzhe Wang, Huaicheng Zhang, Lishi Zuo, and Zhiyong Wu. SongBench: A fine-grained multi-aspect benchmark for song quality assessment. *arXiv preprint arXiv:2604.25937*, 2026. URL https://arxiv.org/abs/2604.25937.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. URL https://ieeexplore.ieee.org/document/10095969/.

Yuanming Yang, Guoqing Ma, Bo Wang, Yuan Zhang, Wei Tang, Chenyi Li, Haoyang Huang, and Nan Duan. DiT-Reward: Generative representations for text-to-image reward modeling. *arXiv preprint arXiv:2606.23626*, 2026. URL https://arxiv.org/abs/2606.23626. Preprint; not peer reviewed.

Jixun Yao, Guobin Ma, Huixin Xue, Huakang Chen, Chunbo Hao, Yuepeng Jiang, Haohe Liu, Ruibin Yuan, Jin Xu, Wei Xue, Hao Liu, and Lei Xie. SongEval: A benchmark dataset for song aesthetics evaluation. *arXiv preprint arXiv:2505.10793*, 2025. URL https://arxiv.org/abs/2505.10793.

Di Zhu and Zixuan Li. MuQ-Eval: An open-source per-sample quality metric for AI music generation evaluation. *arXiv preprint arXiv:2603.22677*, 2026. URL https://arxiv.org/abs/2603.22677.

Haina Zhu, Yizhi Zhou, Hangting Chen, Jianwei Yu, Ziyang Ma, Rongzhi Gu, Yi Luo, Wei Tan, and Xie Chen. MuQ: Self-supervised music representation learning with mel residual vector quantization. *arXiv preprint arXiv:2501.01108*, 2025. URL https://arxiv.org/abs/2501.01108.

Jiabao Zhuang, Changhao Jiang, Hanchen Wang, Jiahao Chen, Zhixiong Yang, Zhenghao Xiang, Yifei Cao, Jiajun Sun, Hui Li, Ming Zhang, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. MuseCritic: Learning multi-aspect song rewards through natural-language aesthetic critiques. *arXiv preprint arXiv:2608.11755*, 2026. URL https://arxiv.org/abs/2608.11755.

Guide to the supplementary material

The supplementary material provides methods and supporting analyses for the paper's scientific questions. The following routes connect each main-text argument to its evidence and scope.

Question in the main text	**Supporting material**
What supports condition discrimination?	Appendices A–C give construction, scoring policies, donor comparisons, input-visibility diagnostics, human quality control, and computation.
How are direct quality and teaching utility measured?	Appendix D.1 specifies the predictors and students. Appendices D.3, D.4, and D.5 establish fitting scopes and calibration; Appendix D.8 defines the supervision interface.
What do common human targets establish?	Appendices D.11 and D.12 report the student listening test and common-target study. Appendix D.13 compares both dependent reannotation rounds.
Can the judgments improve an evaluator?	Appendices D.14 and D.15 detail feedback adaptation, matched controls, and content/listener holdouts.
Why can direct quality and utility differ?	Appendix E contains the counterexample proof, local gradient derivation, and a separate sufficient Brier-risk regime.
Which calculations can be reproduced?	Appendix D.6 maps the numerical comparisons to reproduction packages and states their computational scope.

A Control Validity: Data, Scoring Policies, and Additional Analyses

A.1 Candidate retention and accepted-set composition

Before annotation, each family has 80 items and each canonical direction has 40. Human verification changes both the family sizes and the label proportions. Table 5 gives the initial-evaluation counts; *2+nonbinary* denotes two matching binary votes and one *both*, *neither*, or *unsure* response.

Table 5: Initial control-candidate retention by attribute. Acceptance requires two of three listeners to choose the same condition. Forward/reverse counts refer to opposite predefined edit directions; source/ counterfactual counts refer to the accepted human label. The final columns partition accepted items into unanimous binary votes, split binary votes, and two matching binary votes plus one both/neither/unsure response.

Attribute	Proposed	Accepted (retained)	Forward / reverse	Source / counterfactual	3–0 binary	2–1 binary	2+nonbinary
Genre	80	64 (80.00%)	34 / 30	60 / 4	39	9	16
Instrument	80	54 (67.50%)	28 / 26	43 / 11	33	9	12
Mood	80	67 (83.75%)	34 / 33	55 / 12	39	14	14
Tempo	80	69 (86.25%)	32 / 37	46 / 23	43	17	9
Vocal gender	80	57 (71.25%)	31 / 26	50 / 7	46	6	5
Vocal presence	80	73 (91.25%)	37 / 36	56 / 17	39	8	26
Total	**480**	**384 (80.00%)**	**196 / 188**	**310 / 74**	**239**	**63**	**82**

All 23 generator labels survive the acceptance filter. This is diversity within the CMI-Pref-Pseudo source, not an unseen-generator result.

A.2 Exact attribute-level audio-dependence results

Table 6 reports the matched-minus-donor accuracy gap in percentage points and its prompt-grouped 95% interval. An asterisk marks a lower bound above zero. We report all 36 model–attribute cells.

The ACE-Step probe's prespecified rule required the grouped 95% CI lower bound to exceed 50% for matched-audio raw accuracy and zero for G_D, both overall and in at least four of the six attribute families, including at least one of vocal gender or vocal presence. The overall criteria were met; mood and vocal gender were the only qualifying families, so the four-family requirement was not met.

Table 6: Waveform-dependence gaps across 384 initial accepted items: matched-minus-donor raw accuracy, in percentage points, with conditions fixed. Positive values favor matched audio; brackets are nominal prompt-bootstrap 95% confidence intervals. An asterisk marks an unadjusted lower bound above zero, not passage of the composite family test. TuneJury A5 uses CLAP audio/text, A6 uses MERT audio/CLAP text, and full uses both audio features plus CLAP text; L8 denotes decoder layer 8.

Attribute	ACE-Step v1.5 XL-SFT L8	MuQ-MuLan	CLAP	TuneJury A5	TuneJury A6	TuneJury full
Genre	$+14.06$ $[0.00, 28.13]$	$+40.62^*$ $[26.56, 54.69]$	$+45.31^*$ $[31.25, 57.81]$	$+15.62^*$ $[3.12, 28.12]$	$+17.19^*$ $[4.69, 29.69]$	$+12.50$ $[0.00, 25.00]$
Instrument	$+14.81$ $[0.00, 29.63]$	$+25.93^*$ $[11.11, 40.74]$	$+46.30^*$ $[31.48, 61.11]$	$+5.56$ $[-9.26, 20.37]$	$+18.52^*$ $[5.56, 31.48]$	$+0.00$ $[-11.11, 11.11]$
Mood	$+17.91^*$ $[4.48, 31.34]$	$+17.91^*$ $[4.48, 31.34]$	$+14.93$ $[0.00, 29.85]$	$+2.99$ $[-8.96, 14.93]$	$+10.45$ $[0.00, 20.90]$	$+7.46$ $[-5.97, 20.90]$
Tempo	$+10.14$ $[-4.35, 24.64]$	$+21.74^*$ $[8.70, 34.78]$	$+13.04$ $[0.00, 26.09]$	$+7.25$ $[-5.80, 20.29]$	$+14.49^*$ $[2.90, 26.09]$	$+8.70$ $[-2.90, 20.29]$
Vocal gender	$+36.84^*$ $[22.81, 50.88]$	$+12.28$ $[-3.51, 28.07]$	$+33.33^*$ $[15.79, 50.88]$	-1.75 $[-15.79, 12.28]$	-1.75 $[-15.79, 12.28]$	$+7.02$ $[-5.26, 19.30]$
Vocal presence	-5.48 $[-19.18, 8.22]$	$+12.33^*$ $[1.37, 23.29]$	$+13.70^*$ $[1.37, 26.03]$	$+9.59$ $[-2.74, 21.92]$	$+15.07^*$ $[2.74, 27.40]$	$+2.74$ $[-9.59, 15.07]$

The archived analysis includes family-level matched and donor-audio accuracy, confidence intervals, paired counts, and McNemar tests.

A.3 COMPLETE OVERALL MODEL RESULTS

The exact initial and follow-up model results are reported below. Initial BA has no interval because that interval was not part of the prespecified analysis; the follow-up reports a human-label-stratified interval.

Table 7: Initial control accuracy and waveform dependence on 384 human-accepted items. Accuracy and balanced accuracy (BA) are percentages; matched-minus-donor accuracy gaps are percentage points. Donor audio replaces the recording while keeping conditions fixed; 95% confidence intervals use prompt bootstrap. Always Source selects the unedited condition without audio. TuneJury A5/A6/full use CLAP/MERT/both audio features with CLAP text. Bold marks the highest matched accuracy, BA, and gap, excluding Always Source.

Method	Evidence status	Matched raw accuracy (95% CI)	Matched-audio BA	Donor raw accuracy	Matched - donor raw gap (95% CI), points
ACE-Step v1.5 XL-SFT layer 8	Prespecified evaluation	70.05 (65.36–74.48)	62.42	56.25	+13.80 (+8.07 to +19.53)
MuQ-MuLan	Prespecified control	71.61 (67.19–76.04)	**71.10**	50.00	+21.61 (+15.89 to +27.08)
LAION-CLAP-Music	Post-hoc exploratory	**72.92 (68.23–77.34)**	70.88	46.35	**+26.56 (+20.57 to +32.55)**
TuneJury A5	Post-hoc exploratory	53.39 (48.44–58.33)	53.13	46.61	+6.77 (+1.56 to +11.99)
TuneJury A6	Post-hoc exploratory	58.59 (53.65–63.54)	59.95	46.09	+12.50 (+7.55 to +17.71)
TuneJury full	Post-hoc exploratory	53.91 (48.96–58.85)	54.48	47.40	+6.51 (+1.56 to +11.46)
Always Source	Descriptive	80.73	50.00	80.73	0.00

Table 8: Frozen-policy follow-up on 219 human-accepted items, disjoint from the initial samples and excluding ACE-Step-generated recordings. BA is balanced accuracy; gaps are matched-minus-donor raw accuracy in percentage points, with conditions fixed. Nominal 95% prompt-bootstrap intervals stratify BA by human label. Always Source ignores audio. TuneJury A5/A6/full use CLAP/MERT/both audio features with CLAP text. Bold marks the highest matched accuracy, BA, and gap, excluding Always Source. The coverage extension makes this a sensitivity analysis.

Method	Matched raw accuracy	Matched-audio BA (95% CI)	Donor raw accuracy	Matched - donor raw gap (95% CI), points
ACE-Step v1.5 XL-SFT layer 8	68.49%	**65.20% (56.69–73.27)**	59.36%	+9.13 (+1.37 to +16.89)
MuQ-MuLan	68.04%	64.92% (56.45–72.84)	49.32%	+18.72 (+10.96 to +26.48)
LAION-CLAP-Music	**71.23%**	64.93% (56.46–73.14)	49.77%	**+21.46 (+13.70 to +28.77)**
TuneJury A5	52.51%	47.65% (39.33–56.29)	48.86%	+3.65 (-3.20 to +10.50)
TuneJury A6	51.14%	47.79% (39.30–56.15)	44.75%	+6.39 (+0.46 to +12.33)
TuneJury full	50.68%	49.45% (40.83–58.07)	44.75%	+5.94 (-0.91 to +12.79)
Always Source	81.74%	50.00%	81.74%	0.00

A.4 STATISTICAL PRECISION AND DONOR MATCHING

Every accepted item has a unique prompt. Initial family sizes range from 54 to 73 and follow-up sizes from 30 to 43. An optimistic two-sided exact-binomial benchmark at level 0.05 and 80% power detects accuracies of roughly 67–69% initially and 72–75% in the follow-up, assuming independent items. This does not describe power for balanced accuracy or paired gaps; family-level non-confirmation therefore is not evidence of a true null effect.

Generator matching succeeds for 383/384 initial donors and 213/219 follow-up donors. The six unmatched follow-up trials occur in genre, instrument, and vocal gender. On the matched subset, CLAP's matched-minus-donor gap is +22.1 points, close to the +21.46-point overall result; the unmatched subset is too small ($n = 6$) for an inferential sensitivity claim.

Post-hoc acoustic and clustering sensitivity. We computed full-view mono root-mean-square (RMS) amplitude in decibels relative to full scale (dBFS) and an energy-weighted spectral centroid over nonoverlapping 4,096-sample Hann-windowed frames, using only frozen waveform views.

Table 9: Post-hoc sensitivity of CLAP's matched-minus-donor accuracy gap G_D (percentage points; positive favors matched audio). Restrictions retain paired recordings in the same pooled quartile of root-mean-square amplitude (RMS), spectral centroid, or both. Brackets are 95% confidence intervals from 10,000 item or generator-cluster bootstrap draws. The item stream (seed 20260826) differs from the original analysis, so identical unrestricted point estimates can have different interval endpoints.

Evaluation	Restriction	Retained item pairs	Generator labels	CLAP G_D [item 95% CI]	Generator-cluster 95% CI
Initial	None	384	23	+26.56 [20.57, 32.55]	[20.75, 33.93]
Initial	Same RMS quartile	123	21	+26.02 [15.45, 36.59]	[16.08, 37.63]
Initial	Same centroid quartile	132	21	+21.97 [11.36, 31.82]	[12.62, 30.32]
Initial	Same both quartiles	49	17	+24.49 [8.16, 40.82]	[11.54, 39.29]
Follow-up	None	219	20	+21.46 [14.16, 28.77]	[16.76, 24.37]
Follow-up	Same RMS quartile	75	14	+29.33 [17.33, 41.33]	[22.22, 36.51]
Follow-up	Same centroid quartile	53	17	+15.09 [3.77, 28.30]	[4.00, 24.39]
Follow-up	Same both quartiles	16	9	+25.00 [6.25, 50.00]	[5.00, 60.00]

Matched-minus-donor mean differences are +1.14 dB (95% item-bootstrap CI $[-0.35, 3.11]$) and -23.7 Hz ($[-84.8, 36.1]$) initially, and +0.78 dB ($[-0.33, 2.05]$) and -23.6 Hz ($[-86.0, 39.2]$) in the follow-up. We then pool matched and donor views within each evaluation, form quartiles, and retain pairs in the same bin. Table 9 reports item and generator- cluster bootstrap intervals. The restriction preserves a positive point estimate, but the joint subsets are small. Matching these two summaries cannot rule out timbre, production-quality, or musical-complexity confounding, so G_D remains a waveform-dependence negative control rather than a causal attribute-listening effect.

Conservative follow-up interval sensitivity. For the three declared 219-item aggregate endpoints, a post-hoc Bonferroni family uses two-sided 98.33% intervals. CLAP BA is 64.93% $[54.92, 74.93]$, G_D is +21.46 $[12.33, 30.59]$, and its paired BA difference from full TuneJury is +15.48 $[2.13, 29.24]$. All retain the declared directions. This post-hoc calculation does not supply an alpha-spending rule or convert the follow-up into confirmation. The audit is `donor_acoustic_sensitivity_v1/AUDIT.json`; its input hashes and analysis script are retained in the local experiment archive.

The frozen provenance also permits a model- and label-blind duration audit. The stored waveform views are resampled to 24 kHz, capped at 30 seconds, and at least 10 seconds long; scorers apply their own input policies, including CLAP's 48 kHz resampling. Matched and donor durations are identical for 339/384 initial pairs and 214/219 follow-up pairs; their absolute difference is at most 0.1 seconds for 375/384 and 215/219 pairs, respectively. Channel counts match for 384/384 and 215/219 pairs.

Cohen's h against 50% BA provides a standardized descriptive effect size. For MuQ-MuLan, CLAP, and full TuneJury, respectively, h is 0.436, 0.431, and 0.090 in the initial evaluation and 0.303, 0.303, and -0.011 in the follow-up. These effect sizes are descriptive and do not add new hypothesis tests. The percentage-point scale in Tables 7 and 8 is more direct: Cohen's h is retained only as a standardized transformation, not as deployment accuracy or an additional decision criterion.

The post-hoc extension sensitivity in Table 10 reports both the 195-item pre-extension union and the 219-item final union. BA and model-difference intervals are grouped by prompt and stratified by human label; raw paired-gap intervals are grouped by prompt. Each interval uses 10,000 resamples under its stated scheme. The wider post-hoc 99% sensitivity intervals clear the three model boundaries in both sets. No alpha-spending or adaptive sample-size correction was predeclared, so these are nominal sensitivity intervals rather than repaired confirmatory inference. The extension was model-blind and not repeated to significance, but that safeguard does not retroactively satisfy the original stopping rule.

Table 11 reports the secondary follow-up attribute estimates.

A.5 ACE-Step v1.5 XL-SFT probe training provenance

The exploratory representation diagnostic compared one-based decoder layers 8, 16, 24, and 32 on the same 726-record development benchmark. Each layer used a scale-only, no-intercept RidgeClassifier with $\alpha = 10$. Its input averaged two fixed noise realizations and concatenated temporally pooled statistics from the median and high-quantile trajectory locations. Five-fold out-of-fold evaluation

Table 10: Post-hoc follow-up sensitivity before and after the model-blind coverage extension. Balanced accuracy (BA) is in percent; differences are percentage points. The donor comparison changes audio while fixing conditions. All intervals are nominal prompt-bootstrap confidence intervals; BA measures also stratify by human label. Neither confidence level corrects the deviation from the original stopping rule.

Analysis set	Items	Metric	Estimate	95% CI	99% CI
Pre-extension	195	CLAP matched-audio BA	66.88%	58.04–75.45	55.18–77.95
Pre-extension	195	CLAP matched-minus-donor raw gap	+23.08	+14.87–+31.28	+12.31–+33.33
Pre-extension	195	Paired matched-audio BA difference (CLAP - full TuneJury)	+19.20	+7.95–+30.31	+3.71–+33.93
Final	219	CLAP matched-audio BA	64.93%	56.46–73.14	53.53–75.49
Final	219	CLAP matched-minus-donor raw gap	+21.46	+13.70–+28.77	+11.87–+31.51
Final	219	Paired matched-audio BA difference (CLAP - full TuneJury)	+15.48	+4.50–+26.47	+0.89–+30.22

Table 11: Secondary follow-up CLAP results by attribute. BA is balanced accuracy; gaps are matched-minus-donor raw accuracy in percentage points, holding conditions fixed. Holm correction covers 12 tests: BA above 50% and gap above zero for each attribute. "Both pass" requires both adjusted p-values below 0.05; bold marks the passing result.

Attribute	Accepted items	CLAP matched-audio BA	Raw matched - donor gap, points	Both Holm tests pass?
Genre	43	53.37%	+13.95	No
Instrument	38	59.52%	+26.32	No
Mood	34	75.60%	+23.53	No
Tempo	38	39.58%	+15.79	No
Vocal gender	30	79.37%	+36.67	**Yes**
Vocal presence	36	78.82%	+16.67	No

grouped records by prompt. Layer 8 was selected after inspecting accuracy and the matched-minus-donor gap across the four candidates; no independent layer-selection set or nested cross-validation was used. Subsequent leave-one-attribute-out, permutation, and nuisance checks were post-selection diagnostics. The four-layer comparison therefore introduces selection bias, and its results remain exploratory. The selected layer and ridge penalty were frozen before the new human-verified labels were collected.

The frozen extractor uses the post-block hidden state of decoder layer 8 at trajectory locations $t = 0.401312$ and $t = 0.608645$. These are the median and 0.8 quantile of the model's logistic-normal training-time timestep distribution with $\mu = -0.4$ and $\sigma = 1.0$. Two fixed noise initializations, 20260803 and 20260804, are evaluated with common noise across the source, null, and counterfactual conditions. For each location and initialization, the 2,560-dimensional hidden state is pooled over the temporal axis by its mean and standard deviation. We subtract the counterfactual-minus-null residual from the source-minus-null residual, average the result across the two noise initializations, and concatenate the two trajectory locations and two pooling statistics. The resulting probe input has $2 \times 2 \times 2{,}560 = 10{,}240$ dimensions. The two noise realizations are therefore averaged rather than concatenated.

The probe was trained to distinguish rule-derived source-versus-counterfactual directions on 726 controlled edits, using confidence-derived sample weights; these targets were not human-verification labels. Multiple edit directions can share a source row, leaving 635 unique source rows, 599 unique prompts, and 599 unique recordings across 21 generator labels. Donors match the source generator for 721/726 training records.

The probe-training set shares no prompts or recordings with either later candidate pool. It shares 21 generator labels with the initial evaluation and 19 with the follow-up. This supports sample disjointness and absence of human-label leakage, but not unseen-generator generalization.

A.6 MuQ, CLAP, and TuneJury inference details

MuQ-MuLan uses the difference between the audio embedding's dot-product similarity to the source and counterfactual texts. CLAP uses the cosine margin between released L2-normalized music-audio and text embeddings. Following the released non-fusion inference path, each waveform is decoded to 48 kHz and a candidate-specific, fixed pseudorandom contiguous window of 480,000 samples (ten seconds) is drawn from all feasible start positions within the at-most-30-second waveform. The assignment depends only on the candidate identifier, is shared across the two condition texts, and is independent of human labels and model outcomes. A single window may miss a late-onset voice, instrument, or structural change; evaluating additional windows would require new feature extraction beyond the cached embeddings.

The released TuneJury heads are evaluated in their documented feature order. A5 concatenates the 512-dimensional CLAP audio and text vectors. A6 concatenates a 1,024-dimensional temporally mean-pooled MERT audio vector with the 512-dimensional CLAP text vector. The full head concatenates CLAP audio, MERT audio, and CLAP text into 2,048 dimensions. MERT consumes the at-most-30-second waveform; every TuneJury path that uses CLAP audio receives the same deterministic crop above. Audio-only heads are omitted because the two candidate conditions share one waveform and would therefore have zero condition margin. The empty-prompt fallback is also inapplicable because it would erase the text contrast. No TuneJury head is fitted or recalibrated for this task.

This policy comparison uses unequal supervision and cannot attribute the observed transfer differences to the encoder representation, learned head, prompt-format shift, or calibration. For the external experiment, the primary learned comparator is the released head whose training mixture excludes Music Arena; shared encoder-pretraining overlap remains incompletely auditable.

A.7 External evaluation and resampling sensitivity

The Music Arena accepted set contains 359 unique prompts in 189 anonymized source-user groups. The four exact generator versions contribute 147 `sonauto-v2-2`, 30 `sonauto-v3-preview`, 62 `lyria-3-30s`, and 120 `lyria-3-pro-preview` accepted items. Table 12 reports the frozen aggregate policy results; intervals for BA are human-label stratified and grouped by source-user identifier.

CLAP's paired BA difference is +14.09 points [6.30, 21.91] over the leave-Music-Arena head and +13.57 [5.51, 21.96] over the secondary full head. Descriptively, its matched BA is 63.33% [54.89, 71.92] on Sonauto and 46.26% [36.15, 57.31] on Lyria. This lineage split was not a frozen interaction test and does not authorize a lineage ranking. Attribute BAs range from 47.09% for tempo to 72.38% for vocal gender; all attribute estimates are secondary and none is a family-level confirmation claim.

Frozen zero-margin accounting. Exact score ties remain in the denominator and receive zero correctness credit under the strict positive-margin rule for control judgments. Among the 359 external

Table 12: Frozen control policies on 359 accepted Music Arena items from 189 source users. BA is balanced accuracy (%); the gap is matched-minus-donor raw accuracy in percentage points, holding conditions fixed. Brackets are nominal 95% user-group bootstrap intervals; BA resamples within label classes. "Leave Music Arena" excludes that dataset from head training; full TuneJury is secondary. Exact score ties receive no credit.

Policy	Matched raw accuracy	Matched BA [95% CI]	Donor raw accuracy	Raw matched-minus-donor gap [95% CI]
CLAP	57.38%	57.30% [50.51, 64.36]	35.65%	+21.73 [15.57, 28.15]
TuneJury leave Music Arena	48.47%	43.21% [37.17, 49.85]	49.86%	-1.39 [-5.78, 3.37]
TuneJury full (secondary)	47.35%	43.73% [37.42, 50.44]	43.73%	+3.62 [-1.49, 9.00]

items, matched/donor tie counts are 82/82 for CLAP, 75/70 for leave-Music-Arena TuneJury, and 73/76 for full TuneJury. The initial and follow-up evaluations have no exact ties. These counts are verified directly from the stored predictions. The aggregate results therefore include each frozen policy's failure to express a directional preference. The input-level diagnosis below identifies the preprocessing source of the CLAP ties and the shared text-vector collapse in the learned heads.

Input visibility and descriptive scoring sensitivity. We trace the external ties using the saved conditions, local RoBERTa tokenizer files, and cached CLAP text vectors, without checkpoint inference. The released CLAP wrapper pads/truncates text to 77 tokens. In each of the 359 pairs, the two condition strings and their untruncated token sequences differ, but 82 pairs become identical after truncation; the same 82 have exactly equal cached text vectors. Of these, 65 have an original-condition human label and 17 a counterfactual label. In total, 165 pairs exceed 77 tokens. Thus the condition contrast is absent at the scorer's input on 22.84% of accepted items, irrespective of waveform quality.

All CLAP matched/donor ties coincide with these 82 pairs. All exact TuneJury ties also belong to this set. The leave-MA head has 7/12 nonzero matched/donor margins on identical text inputs, and the full head has 9/6; every such absolute residual is at most 1.20×10^{-7}. Since each same-audio feature pair is identical, these residuals cannot encode the changed condition. Frozen strict-positive scores are preserved, including these small numerical signs; the diagnostic neither refits the heads nor selects a new threshold.

Table 13: Post-hoc tie sensitivity on 359 Music Arena control items (all entries in percent). BA is balanced accuracy; coverage is the nonzero-margin fraction. Strict gives exact ties no credit, Half gives half credit, and Nonzero retains each policy's directional subset. Visible uses the same 277 token-distinguishable items for all policies. "Leave MA" denotes TuneJury trained without Music Arena. These point estimates add no confirmatory tests.

Policy	Coverage	Strict BA	Half BA	Nonzero BA	Visible BA
CLAP	77.16	57.30	68.27	73.44	73.44
TuneJury leave MA	79.11	43.21	53.54	54.54	54.71
TuneJury full	79.67	43.73	53.46	54.49	54.53

Half-credit scoring preserves the descriptive CLAP–TuneJury ordering (Table 13). The common visible subset has 210 original and 67 counterfactual labels; both learned heads also give nonzero margins throughout that subset. Its CLAP advantage is therefore not explained solely by truncation or exact-tie penalties. This restricted result is conditional on surviving preprocessing, not a replacement population for the frozen external test. Strict BA combines directional coverage with correctness; a low value need not mean systematic reversal of the attribute. A reusable preflight check should ensure that the changed condition survives text processing before interpreting waveform-dependent control scores.

Table 14: Music Arena control-input collapse and directional coverage. Collapsed counts condition pairs made identical by CLAP's 77-token text limit. The last three columns give the percentage of matched-audio scores with nonzero margins. Leave MA/full denote TuneJury heads trained without/ with Music Arena. Tiny numerical head residuals can produce nonzero margins on collapsed inputs; they do not recover the missing condition contrast.

Attribute	Items	Collapsed	CLAP (%)	Leave MA (%)	Full (%)
genre	61	19	68.85	70.49	68.85
instrument	51	5	90.20	94.12	92.16
mood	64	15	76.56	76.56	78.12
tempo	63	17	73.02	77.78	77.78
vocal gender	58	9	84.48	84.48	89.66
vocal presence	62	17	72.58	74.19	74.19

Table 14 breaks down collapse and directional coverage by attribute. The separately hashed diagnostic records counts, per-label margins, per-attribute coverage, token visibility and feature equality in `external_tie_diagnostic_20260912/AUDIT.json`. It leaves the frozen labels, predictions, intervals and decision rule intact.

Table 15: Initial control-label reliability by attribute. Fleiss' κ uses source, edited, both, neither, and unsure responses; exact agreement averages the three listener pairs. Acceptance requires two matching binary votes, and unanimous share is the fraction of accepted items with three matching binary votes. These are descriptive agreement measures.

Attribute	Candidates	Five-category κ	Exact agreement	Accepted	Unanimous share of accepted
Genre	80	0.271	65.0%	64	60.9%
Instrument	80	0.472	66.7%	54	61.1%
Mood	80	0.313	63.3%	67	58.2%
Tempo	80	0.377	65.8%	69	62.3%
Vocal gender	80	0.695	82.9%	57	80.7%
Vocal presence	80	0.343	64.6%	73	53.4%

Post-hoc whole-user bootstrap sensitivity. The frozen BA routine resamples user groups independently within each human-label class. Of 189 users, 26 contribute both labels, so that routine does not preserve their cross-class dependence. A separate 10,000-draw sensitivity (seed 20260912) samples whole users, retains all their items together, and recomputes equal-class BA in each draw. CLAP BA is 57.30% $[49.93, 64.81]$; the donor gap is +21.73 points $[15.72, 27.95]$; and the paired BA difference from leave-Music-Arena TuneJury is +14.09 $[6.04, 22.20]$. Every draw contains both label classes. The relative policy and donor contrasts retain positive intervals, but the BA interval includes 50%, so the full three-part claim is sensitive to clustering. This post-hoc diagnostic leaves the frozen results intact and does not repair or replace their decision rule. Its inputs and results are recorded in `experiments/generative_reward/external_whole_user_sensitivity_20260912/AUDIT.json`.

B Control Annotation and Quality Control

B.1 Packet quality control

This audit excludes names and free-text notes. Each of three packets contains the same 480 base candidates plus 24 orientation-reversed hidden repeats and is divided among four contributors; the 12 contributors are disjoint across packets. Repeat consistency is 95.83%, 95.83%, and 100%. Every candidate therefore has three distinct base judgments, while repeats check packet quality and do not enter evaluation targets. Five-category Fleiss' κ across the base judgments is 0.431; mean pairwise exact agreement is 68.06%, compared with 43.83% expected from the marginal response frequencies.

Across accepted items, 1,007/1,152 votes (87.41%) match the in-sample majority label. Because the reference label is built from the same votes, this is an optimistic reference-reliability diagnostic rather than an independent annotator ceiling.

The family audit in Table 15 shows that ambiguity is heterogeneous. Genre and mood have the lowest five-category κ, while vocal gender has both the highest κ and the largest unanimous share. Tempo is intermediate rather than uniformly easier than mood. The within-family estimates have only 80 candidates each and are diagnostic.

A post-hoc sensitivity analysis splits accepted labels without changing them or adding adjudication. Table 16 reports the paired difference in matched-audio BA (CLAP minus full TuneJury); intervals use 10,000 label-stratified resamples. Both evaluations have larger point estimates on unanimous items and intervals crossing zero on non-unanimous items. This descriptive pattern does not establish a between-subgroup difference or identify label ambiguity as its cause; the aggregate analysis retains the prespecified majority rule.

Unanimous-only performance is not a label-noise upper bound: unanimity selects a different subset using the same votes that define its labels. We do not report family-level unanimous BA because some resulting class cells are absent or extremely small; for example, the follow-up instrument subset contains no unanimous counterfactual label. Independent expert adjudication would provide a separate reference for assessing label accuracy.

Table 16: Post-hoc control sensitivity to vote unanimity. Balanced accuracy (BA) is in percent; paired CLAP-minus-full-TuneJury differences are percentage points with nominal 95% label-stratified bootstrap intervals. Source/counterfactual counts describe human labels. Unanimous items have three matching binary votes. The rows do not test a difference between subgroups.

Evaluation and vote subset	**Items (source / counterfactual)**	**CLAP BA**	**Full TuneJury BA**	**Paired BA difference (95% CI), points**
Initial, unanimous	239 (205 / 34)	76.5%	51.6%	+24.9 (+12.4 to +36.7)
Initial, non-unanimous	145 (105 / 40)	64.8%	55.5%	+9.2 (-3.3 to +22.3)
Follow-up, unanimous	156 (132 / 24)	68.8%	45.6%	+23.1 (+10.0 to +36.2)
Follow-up, non-unanimous	63 (47 / 16)	59.1%	53.6%	+5.5 (-13.2 to +23.4)

B.2 Follow-up coverage-deviation audit

The audit retains the full follow-up sequence, not only the final passing union. The first round accepted 187/240 items and had three perfect repeat checks. The planned supplement reannotated all 31 unresolved genre/instrument candidates from that original pool, using only their three new votes and preserving the 187 accepted labels. It had packet repeat rates of 100%, 80%, and 100% and added eight majorities, but the 195-item union still missed the original genre and instrument thresholds. No model output had been produced.

The separately specified extension used 32 new candidates and five hidden repeats per packet. All three packets returned 37/37 rows, and every repeat check was 5/5. Twenty-four candidates gained a majority: 14 unanimous source, 3 unanimous counterfactual, 4 with two source votes and one non-binary response, 1 with a two-to-one source majority, and 2 with a two-to-one counterfactual majority. The 219-item union contains 179 source and 40 counterfactual labels. Names and free-text notes are excluded from all follow-up analyses.

Satisfying the criterion with the extended union does not retroactively satisfy the original stopping rule.

B.3 External-confirmation gate audit

Each Music Arena packet returned all 402 requested rows: 384 base items and 18 orientation-reversed repeats. All 18 repeat pairs agree in every packet. Public validation passes before private A/B orientation is opened. Anonymous aggregation retains 359 binary majorities, and every frozen coverage threshold passes without a supplement or extension.

The accepted human labels are 275 original and 84 counterfactual. Reports exclude annotator identities and free-text notes. One empty administrative annotator-ID cell in packet 2 was restored from that packet's sole stable nonempty ID under a predeclared mechanical normalization; no task ID, response, confidence, note, or row order changed.

C Reproducing the Control Evaluations

For the initial and same-source follow-up evaluations, the donor rule is as follows. For target i, define $E_i = \{j : p_j \neq p_i, x_j \neq x_i, a_j \neq a_i\}$, where p, x, a, and g denote prompt, recording, attribute, and generator identifiers. Define $G_i = \{j \in E_i : g_j = g_i\}$, and set $D_i = G_i$ if G_i is nonempty and $D_i = E_i$ otherwise. The donor is $j^* = \arg\min_{j \in D_i} r(i, j)$, where r is the fixed pseudorandom ordering of target–candidate identifier pairs implemented in the archived code. Construction stops if no eligible donor exists. The external Music Arena rule additionally excludes the same source user and item; its fallback after exact generator matching is restricted to the same generator lineage, with failure if neither pool is available. Neither rule relaxes the prompt, recording, or attribute exclusions, and neither ordering uses human labels, model features, or model scores.

The full local control-evaluation archive separates immutable source data, annotation responses, model predictions, and derived analyses. Candidate construction records the attribute edit, generator provenance, and donor assignment for every item. This internal archive retains randomized task materials and anonymous responses, with restricted access to the mapping between display order and condition identity. These materials and raw responses are excluded from the prepared statistical supplement. Any later release follows the timing, approval, and rights conditions stated in the ethics and reproducibility statements. Model and checkpoint versions are recorded together with the preprocessing policies used for each representation. The fixed scoring and preprocessing policies are specified in Appendix A.6.

Reported tables, figures, agreement statistics, and confidence intervals are generated from archived item-level records by scripted workflows. A single-thread CPU replay in the archived software environment, consuming stored model features, reproduced all 219 follow-up predictions and the resulting aggregate statistics byte-for-byte. It does not rerun GPU feature extraction and is therefore a deterministic statistical replay, not an independent end-to-end reproduction.

For the external confirmation, the pre-score audit freezes 359 accepted scoring rows and all 384 feature candidates. A separate cache audit verifies exact candidate coverage, prompt and audio hashes, deterministic CLAP crop seeds, tensor shapes and finite values, rank-partition completion, and every frozen code and checkpoint hash. Extraction used Python 3.12.3, PyTorch/torchaudio 2.6.0, transformers 4.44.0, NumPy 1.26.0, and librosa 0.11.0 on eight accelerator devices. Control evaluation did not use the official CMI test; its one-time use for direct preference and student evaluation is specified in Appendix D.9.

D Preference Prediction, Student Supervision, and Feedback Adaptation

The methods below define the compared systems before the evidence summary. Data-use boundaries and the isolated follow-up are in Appendices D.3 and D.4; executable statistical entries and their scope are in Appendix D.6.

D.1 Teacher and student specifications

Grouping and analysis units. Development votes sharing a prompt identifier are grouped together. Groups are joined transitively when candidate or reference audio has the same file-content hash; the resulting components define the splits and resampling units. Pseudo-source components similarly connect prompt identifiers through candidate-audio hashes. T7 and T29 retain those source-component keys after pair selection. Development NLL weights vote rows equally, whereas T7 and T29 NLL weight scored pairs equally after averaging their valid votes. Contrasts use unrounded losses; displayed values are rounded independently.

Teacher inputs and calibration. With logistic link $\sigma(t) = (1+e^{-t})^{-1}$, no-intercept Bradley–Terry logits are antisymmetric under candidate swaps, so swapped probabilities sum to one (Bradley & Terry, 1952). Table 17 defines the preference scorers used as teachers. The 10D teacher uses four AudioBox outputs (Tjandra et al., 2025): CE (Content Enjoyment), CU (Content Usefulness), PC (Production Complexity), and PQ (Production Quality), each summarized by its mean and standard deviation. MuQ-Eval's MI denotes Musical Impression and is used as a musicality proxy here. For an output j, let $d_j = s_j(x_A) - s_j(x_B)$ and let $r_j = \sqrt{|F|^{-1} \sum_{i \in F} d_{ij}^2}$ on the teacher fitting rows F; elementwise scaling requires finite $r_j > 10^{-12}$. The 4D, 10D, and 22D teachers fit $p_A = \sigma(\boldsymbol{\beta}^\top(\mathbf{d}/\mathbf{r}))$ using L2-regularized logistic regression with $C = 1$, no intercept or class weighting, and L-BFGS with a 10,000-iteration cap. AudioBox instead fits one coefficient to $\frac{1}{2}(d_{\mathrm{CE}}/r_{\mathrm{CE}} + d_{\mathrm{PQ}}/r_{\mathrm{PQ}})$. For development AudioBox probabilities, the scalar coefficient and RMS scales are fitted separately on each outer training fold. All five coefficients are positive, preserving the original 73.46% accuracy. The full-development scalar fit supplies pseudo-labels and independent predictions. External 4D's Music Arena predictions apply its archived full-development calibrator to the same 3,660 battles and cached inputs used by the other scorers, without refitting or selection on Music Arena labels. The fitting scope follows Table 21: direct development excludes the evaluation fold, historical pseudo teachers use all development labels, and T30 rebuilds calibration within each inner or outer fitting

scope. AudioBox scores non-overlapping 10-second windows at 16 kHz mono, zero-padding and masking the final partial window. Track means and population standard deviations weight windows by their valid audio duration. MuQ-Eval uses the A1 checkpoint at 24 kHz mono. We average its MI predictions over distinct 10-second start, center, and end crops; clips shorter than 10 seconds are zero-padded and evaluated once.

Table 17: Inputs to preference-teacher calibrators. D counts score inputs, not independent evaluators. AudioBox CE/CU/PC/PQ denote Content Enjoyment/Content Usefulness/Production Complexity/Production Quality; MuQ-Eval MI denotes Musical Impression. Means and standard deviations summarize windows within a track. RMS denotes root-mean-square scaling of pairwise score differences.

Teacher	Inputs and aggregation before calibration
AudioBox	CE and PQ means, separately RMS-scaled and averaged into one scalar margin.
External 4D	TuneJury reward; AudioBox CE mean; AudioBox PQ mean; MuQ-Eval MI.
10D	TuneJury reward; AudioBox CE, CU, PC, and PQ means and their four window standard deviations; MuQ-Eval MI.
22D	All 10D inputs; SongEval Coherence, Musicality, Memorability, Clarity, Naturalness; SongBench Melody, Arrangement, Musicality, Vocal, Instrumental, Mixing, Structure.

For preference prediction, the TuneJury input uses the leave-Music-Arena checkpoint with a fixed 512D zero text vector. The External-4D gold gate treats TuneJury, AudioBox, and MuQ-Eval as three families: their raw directions are respectively $\mathrm{sign}(d_{\mathrm{TJ}})$, $\mathrm{sign}(\frac{1}{2}(d_{\mathrm{CE}}/r_{\mathrm{CE}} + d_{\mathrm{PQ}}/r_{\mathrm{PQ}}))$, and $\mathrm{sign}(d_{\mathrm{MI}})$. All three must be nonzero and agree with the calibrated hard label; selection, target, weight, and budget rules are detailed in Appendix D.8 and, for T30, Appendix D.4.

Student representations and initialization. All students use cached, frozen audio encoders and native frozen 512D MuQ-MuLan caption/lyrics embeddings. MuQ audio is decoded as 24 kHz mono and pooled over the encoder's 10-second clips, extending a final partial clip with audio from the track's start: frame means are averaged over clips. For Q0, each frame is layer-normalized across its 1,024 coordinates before pooling, without fitted affine parameters; LF0 pools the raw final-layer states. The MERT student configuration (M0) pools raw layer-24 states by frame count over the full 24 kHz track, using 300-second chunks and merging a final tail shorter than one second. Table 18 specifies the paths from these caches.

For Q0's twelve pooled vectors $\mathbf{h}_1, \ldots, \mathbf{h}_{12} \in \mathbb{R}^{1024}$, fusion uses the Gaussian error linear unit (GELU) activation and produces

$$\mathbf{h}_{12} + W_{\mathrm{up}}\ \mathrm{GELU}\big(W_{\mathrm{down}}[\mathbf{h}_1; \cdots; \mathbf{h}_{12}] + \mathbf{b}_{\mathrm{down}}\big) + \mathbf{b}_{\mathrm{up}},$$

where W_{down} maps $12{,}288 \to 64$ and W_{up} maps $64 \to 1{,}024$. The fusion has 853,056 trainable parameters, including biases. Initial Q0 training reloads the same stored random-down/zero-up state, so the initial fused output equals $\mathbf{h}_{12}$; the fusion has no aesthetic pretraining. The common adapters and heads are freshly initialized with the matched fit seed. Subsequent feedback adaptation instead starts from the trained checkpoints (Appendices D.14 and D.15).

Table 18: Student audio paths and trainable modules. Q0 is the MuQ layer-fusion student; LF0 uses only MuQ's final layer; M0 is the MERT configuration. D denotes vector dimension. All audio encoders are frozen; the common scorer trains condition/audio adapters and Musicality/Alignment heads. Reference audio follows the same audio path.

Student	Path before the common scorer	Trainable components
Q0	Twelve normalized MuQ layers → rank-64 residual fusion → frozen MuQ-MuLan audio projection → 512D unit vector.	Fusion; common adapters and task heads.
LF0	Raw final MuQ layer (1024D) → the same frozen MuQ-MuLan audio projection → 512D unit vector.	Common adapters and task heads; no layer fusion.
MERT M0	Raw final MERT-v1-330M layer (1024D), passed directly to the common scorer.	Common adapters and task heads; no layer fusion.

The frozen MuQ-MuLan projection consists of affine maps $1024 \to 768 \to 512$ without an intervening activation, followed by L2 normalization. The common scorer separately maps candidate audio, caption, lyrics, and reference audio to 256D and L2-normalizes each, using $\epsilon = 10^{-8}$. The Musicality head is an audio-only $256 \to 256 \to 64 \to 1$ multilayer perceptron (MLP). The Alignment head takes the elementwise product and inner product of candidate audio with each of the three conditions, plus three presence flags: $3(256 + 1) + 3 = 774$ inputs, followed by $774 \to 256 \to 64 \to 1$. Both heads use GELU and dropout 0.1 after each hidden layer. Absent lyrics/reference vectors are masked to zero; the caption flag is one. Shared heads score the two candidates, producing task logits $\ell_t = s_t(x_A, c) - s_t(x_B, c)$ and probabilities $\sigma(\ell_t)$ for task t, where c collects the conditions and presence flags; Musicality ignores c.

Matched training and checkpoint selection. The mixed warm-up setting (G25) uses a 0.25 pseudo-loss multiplier; the hard-target setting is denoted HARD. Table 19 gives the common initial-training settings for Q0/G25, LF0, and MERT M0 comparisons unless explicitly ablated. Binary cross-entropy (BCE) takes a logit as input: $B(\ell, y) = \log(1 + e^{\ell}) - y\ell$, evaluated by the numerically stable logits routine. With each task loss averaged over its human batch, the human loss is $L_h = 0.25B_{\text{Alignment}} + 0.75B_{\text{Musicality}}$. The mixed loss is $L_h + 0.25 \ \text{mean}_i[w_i B(\ell_i, \hat{y}_i)]$; pseudo pairs supervise only Musicality. The mean divides by the number of pseudo examples in the batch, not by the sum of weights. ZERO changes the pseudo multiplier to zero while replaying the matched schedule and operations. The swap average in Table 19 satisfies $B(-\ell, 1 - y) = B(\ell, y)$: it gives equal weight to the two orientations of the same computed logit, without doubling the number of observations.

Table 19: Common initial student-training settings. n_h and n_p count human and pseudo-labeled training rows; BCE is binary cross-entropy, ℓ a preference logit, and y its target. Early stopping applies only to inner schedule selection; final fits replay the chosen epochs. Mixed epochs combine human and teacher supervision, followed by human-only epochs.

Setting	Value or rule
Optimizer	AdamW; reset optimizer state before human-only fine-tuning.
Batch and gradient	Batch size 256; global gradient-norm clipping at 1.0. Each mixed step uses one human batch and one pseudo batch.
Mixed epoch	$\max(\lceil n_h/256 \rceil, \lceil n_p/256 \rceil)$ updates; deterministically reshuffle and cycle each stream to fill its batches.
Swap augmentation	For each human task, equally average BCE for (ℓ, y) and $(-\ell, 1 - y)$.
Inner search	Learning rates $\{10^{-3}, 3 \times 10^{-4}, 10^{-4}\}$; weight decays $\{0, 10^{-4}\}$; minimize human joint validation BCE.
Stopping per phase	Maximum 120 epochs; stop after at least 12 epochs when 12 epochs fail to improve validation loss by more than 10^{-4}. Retain the best checkpoint, whose epoch can be earlier than 12.
Final schedule	Reload initialization and refit on outer training rows for the selected mixed and human-only epoch counts. Original schedules: Table 27; fresh T30 schedules: Table 23.
Fit seeds	Seed bases 260114, 260214, 260314, 260414, 260514; add the outer-fold index (0–4).

The inner search restores its best mixed-phase checkpoint before starting the human-only phase and selects the learning-rate/weight-decay trial by the latter phase's best validation loss, breaking ties by ascending learning rate and weight decay. Original LF0 and MERT comparisons replay the original Q0 fold schedules under separately frozen protocols. T30 instead uses the first seed's inner External-4D search to choose a fresh schedule shared by every arm and seed within that fold. It compares teachers through this common interface, without teacher-specific schedule optimization. Prediction averaging follows Appendix D.8; seed models are not additional human samples.

D.2 Evidence status and decision hierarchy

Table 20 separates the populations and evidence used by the main argument. "Predeclared" and "frozen" refer to locally archived, hash-bound protocols fixed before the relevant stage outcomes. This timing does not make reused development votes independent evidence. Original development student

comparisons are conditional diagnostics because their external teachers use the full development labels (Appendix D.3); predeclaration and student fold separation do not establish end-to-end OOF isolation. T30 separately supplies a post-hoc, fold-scoped Q0 comparison (Appendix D.4).

Table 20: Preference-evaluation evidence and permitted interpretations. OOF means out-of-fold prediction; NLL is negative log likelihood; Q0 is the MuQ layer-fusion student; QC is quality control. T labels identify archived experiments, and R2/R3 identify successive reannotations of the same pool. Development and reannotation rows reuse data and are not additive samples or independent replications.

Layer	Population	Role
Original development	3,527 votes	Direct OOF evaluation; student diagnostics conditional on full-development teachers.
T30 follow-up	Same 3,527 votes	Fully fold-scoped Q0 teacher–student evaluation; post-hoc, fixed configurations, not a new independent cohort.
T31 calibration	Same 3,527 votes	Symmetric training-side temperature transfer; four-test adjusted comparisons, conditional on fitted systems.
T7	1,200 pairs; 1,195 with ≥ 1 vote; 1,134 with three	Prospective sample-disjoint human confirmation of student NLL contrasts.
Music Arena	3,660 decisive battles; 1,165 users	Frozen zero-refit external ranking evaluation under user-cluster bootstrap.
Official CMI	500 pairs	One-shot, predeclared, no-refit direct and student evaluation.
T29 primary	2,400 selected; 1,000 QC-eligible; 999 evaluated	Prospective same-pair human test of direct/teaching order and embedded multi-aspect construct audit.
T29 R2	Same 2,400 selected; 2,200 QC-eligible; 2,197 evaluated	Post-hoc supplier remediation using unchanged numerical gates and frozen predictions; does not replace the original primary.
T29 R3	Same 2,400 selected; 2,400 QC-eligible; 2,397 evaluated	Second post-hoc reannotation; complete latest returns replace three designated slots; 23/24 people pass.
T32 adaptation	Same 2,397 pairs; 9,020 votes	Post-hoc preference-scorer adaptation with five new component folds; no new cohort or listening.
T33 listener holdout	1,916 development and 473 test pairs	Supplementary post-hoc refit; 17 people supply feedback and six supply test judgments on disjoint sources.

Multiplicity and effect scale. Confirmatory claims follow their original decision rules: every external, official, or original T29 co-primary component must pass its specified boundary. The original T29 secondary families use Holm adjustment for seven Q0 contrasts and seven fixed human-to-output tests. The follow-up control-attribute tests also use Holm adjustment, but remain sensitivity analyses because the original stopping rule was not met. T29 reannotation comparisons remain post hoc under their within-version adjustments. The post-hoc T31 follow-up separately applies Holm correction to four calibrated-student-versus-calibrated-ZERO comparisons; its 95% intervals remain nominal (Appendix D.5). Each of the two post-hoc feedback analyses adjusts its six NLL contrasts as one Holm family; their intervals condition on fitted predictions (Appendices D.14 and D.15). The many development, ablation, and slice intervals are nominal 95% exploratory intervals and are not jointly multiplicity-adjusted. Accuracy effects are reported in percentage points, proper scores in absolute NLL, and Table 28 multiplies utility by 10^3 only for readability; unscaled NLL appears beside it. The paper reports estimates and intervals rather than treating null inclusion as proof of no effect.

Scope and development provenance. Q0 and its content-connected folds were fixed inputs to the multi-teacher protocol. Earlier architecture screens used to develop Q0 are outside this study's confirmatory evidence and are not counted as replications.

D.3 TEACHER–STUDENT FITTING SCOPES AND LABEL REUSE

Original implementation. The direct External-4D continuation gate and 10D comparisons use calibration fitted outside the evaluation fold; 22D additionally nests family selection inside each outer training set. After the direct gate, the original protocol explicitly permits a full-data teacher refit. In `select_t1.py`, all 3,527 human labels and feature rows determine RMS scales and the External-4D logistic calibrator. `prepare_t6_control_targets.py` similarly fits the AudioBox calibrator on all development labels. The 10D and 22D target builders `prepare_t13b_targets.py` and `prepare_t15_targets.py` use all development labels and then generate one pseudo target artifact per teacher. These files are under `multi_teacher_relabel/`; the source-bound inspection is archived in `DEVELOPMENT_DATA_USE_AUDIT_20260912.json`.

In `train_t2.py`, pseudo targets, weights, and selected positions are assembled before the student outer-fold loop and supplied to every fold. Later Q0, LF0, and MERT external-teacher comparisons reuse the corresponding target artifacts. Their outer evaluation labels are excluded from direct human-loss optimization, but the following indirect path remains:

$$\begin{aligned}&\text{outer evaluation labels} \longrightarrow \text{full-development teacher calibration}\\ &\longrightarrow \text{pseudo probabilities, selection, targets, and weights} \longrightarrow \text{student training.}\end{aligned} \tag{6}$$

Q0 self-targets are generated with fold-specific Q0 models, unlike the full-data external calibrators. This distinction does not make contrasts with the external pipelines end-to-end isolated, nor does it remove the common schedule's selection history. Audio-disjoint pseudo data cut content overlap, not the label path in Equation 6.

Interpretation of the original student estimates. The original development student tables, gains against ZERO, rank maps, cross-student comparisons, and related diagnostics describe pipelines conditional on teachers already informed by the complete development set. Their paired intervals summarize loss variation for fixed predictions; they do not estimate the magnitude of the data-reuse bias or remove it. Student-fold exclusion and pre-outcome freezing alone do not establish end-to-end OOF isolation of the combined pipeline.

Independent evaluations and complete fold isolation. Direct development OOF scores use their own fold-specific calibrators and must not be conflated with the subsequent full-data pseudo-label refits. The T7, Music Arena, official, and original T29 targets were not used to fit those teachers or students; their population-specific estimates remain the evidence for generalization of the frozen systems. This does not assert absence of upstream pretraining overlap or make overlapping human cohorts independent replications. Both reannotation rounds use reworked judgments on the original selected pool and remain post hoc. Their labels were not used to fit or recalibrate the models evaluated in those frozen assessments. The subsequent feedback-adaptation experiment uses the latest aggregate labels under separate outer-fold exclusion (Appendix D.14); it does not replace these estimates.

The fully fold-isolated comparison (Appendix D.4) evaluates fixed Q0 configurations: every teacher transformation uses the matching fit scope, and pseudo selection, targets, weights, and inner schedule selection are rebuilt. The original LF0/MERT and Q0 development diagnostics retain their full-development teacher fits. Official outcomes are not used to tune or select the fold-isolated pipelines.

Separate reuse for feedback adaptation. The feedback experiment starts from the original ZERO and 10D Q0 checkpoints, not the T30 refits. New source-component folds isolate its fitting, schedule selection, temperature fitting, and evaluation labels. The new pool is disjoint from pseudo-training components and original development audio hashes. Original teacher fitting on all development labels therefore does not supply a path from these new held-out labels into the starting models. Prior inspection of the new cohort's frozen-model results still informs the experiment's choice, so the new results are post hoc. T7 overlaps this pool and is not a fresh post-adaptation test; the official test is not reused. Historical independent assessments and new held-out adaptation outcomes answer separate questions.

The supplementary person-and-content holdout independently refits these same pre-feedback starting models using new judgments from only 17 of the 23 accepted annotators (Appendix D.15). Its six test annotators contribute no new fitting, schedule-selection, or temperature-fitting labels; test source components are also excluded. Original human-label replay remains restricted to each starting

checkpoint's original training scope. No adapted weights, selected schedules, temperatures, or all-person targets are reused from the preceding adaptation. The split remains post hoc because the complete cohort and earlier results were already known.

Table 21: Which labels can enter fitting in each preference evaluation. OOF means out-of-fold prediction. Feedback adaptation reuses the common-target annotation pool under new fitting scopes; a new split does not make the pool an independent cohort. "Old labels" are the starting models' original human-training labels.

Evaluation	Teacher calibration	Student fitting	Evaluation labels / interpretation
Direct development	Outer training labels; inner family selection	None	Held-out outer fold; direct OOF
Original student development	All development labels	Fold training labels and teacher-derived pseudo targets	Also used by teacher; conditional diagnostic
Complete fold isolation	Outer train only; inner fit only during tuning	Matching fold labels and rebuilt pseudo supervision	Held out throughout; post-hoc, fixed configurations
Frozen human assessments	Development labels only	Frozen models, where tested	Assessment labels unused by these models
Same-pool reannotation	Unchanged	Unchanged	Reworked judgments; post-hoc, dependent cohort
Feedback adaptation	No teacher refit	Old labels plus new outer-train feedback	New evaluation fold excluded; post-hoc adaptation
Person/content holdout	No teacher refit	Old labels plus 17-person feedback on development sources	Six excluded annotators on disjoint sources; post-hoc supplement

D.4 Fully fold-scoped teacher–student follow-up

Purpose and status. T30 addresses the label path in Equation 6 for Q0. It uses the existing 3,527 development votes in 2,551 connected components, the original five outer folds and inner partitions, and existing frozen teacher scores and student features. No new listening or official-test evaluation occurs. AudioBox, External 4D, 10D, and 22D are fixed configurations, each compared with the same ZERO control. The protocol, source code, input hashes, software versions, seeds, and contrasts were frozen before this rerun's teacher fitting. This is a post-hoc follow-up conditional on previously chosen representations and Q0 architecture, not an independent confirmation or a replay of all historical architecture and feature-family selection.

Complete fitting path. For outer fold k, let $H_{-k} = F_k \cup V_k$ denote the outer training votes, partitioned into inner fit and validation votes, and let H_k be the outer evaluation votes. RMS scales and no-intercept logistic calibrators (fixed $C = 1$) are first fitted on F_k only. AudioBox averages separately RMS-scaled CE and PQ margins before fitting its single coefficient; other teachers use their fixed 4-, 10-, or 22-column margin matrices. Their inner-scope pseudo targets, confidence weights, and selectors are rebuilt from these calibrators. The External-4D Q0 anchor trains on F_k and its inner pseudo supervision; only V_k selects a common training schedule from the original three learning rates and two weight decays, with the original two-stage early-stopping rule. The first seed base, 260114, selects the schedule; all five arms and seed bases 260114, 260214, 260314, 260414, and 260514 share that fold's selected recipe. Actual training seeds add k.

Teachers are then refitted on H_{-k}, their outer-scope pseudo supervision is rebuilt, and each student starts again from Q0 initialization using H_{-k} and that supervision. The resulting model predicts only H_k. Thus the data flow is

$$\begin{aligned} F_k &\longrightarrow \text{inner teacher and pseudo supervision} \longrightarrow \text{inner student}, \\ V_k &\longrightarrow \text{common schedule selection}, \\ H_{-k} &\longrightarrow \text{refitted teacher and pseudo supervision} \longrightarrow \text{refitted student} \longrightarrow \widehat{p}(H_k). \end{aligned} \tag{7}$$

No H_k labels enter these fits or schedule choices. Direct teacher probabilities for H_k also come only from its H_{-k} calibrator. There are 30 inner schedule candidates and 125 final student fits. The common External-4D schedule defines the comparison interface; it is not a claim that each teacher is individually optimized.

Matching supervision and compute within each scope. Only the original 61,271 pseudo-train pairs are eligible. Each fit scope constructs its own External-4D gold pool using confidence at least 0.90 and agreement with all three nonzero raw family directions. Its common budget is the smaller of

11,631 and that scope's eligible gold count. External 4D selects the most confident eligible gold pairs; other teachers select their own most confident pseudo-train pairs, with comparison-ID tie-breaking. This yields outer budgets of 11,179–11,631, independently of inner budgets (Table 23). We do not borrow the old full-development gold mask or use an outer refit to determine an inner budget. All teachers use hard targets, weight $2|p-0.5|$, and pseudo multiplier 0.25. ZERO copies the External-4D pool and executes the same batching and forward operations with multiplier zero. Optimizers are reset before human-only fine-tuning. Within each outer fold, all arms share the human rows, initial state, seed, pair budget, epoch counts, and optimizer recipe.

Evaluation and verification. Each vote's student prediction averages the five seeds from its held-out fold; models from the other four folds, which trained on that vote, never enter this average. We report musicality NLL, Brier, accuracy, alignment NLL, and per-seed metrics. The numerical evaluation clips probabilities to $[10^{-6}, 1-10^{-6}]$. The seven fixed student NLL contrasts and three direct contrasts use 10,000 paired component resamples within outer fold. Their unadjusted percentile 95% intervals condition on fitted predictions and do not capture all uncertainty from resampling and refitting overlapping training sets. All 125 fits must be present before the aggregate evaluation; no candidate is selected on intermediate outer-fold results.

Fourteen targeted tests include poisoning non-fit labels/features, checking unchanged teacher targets and actual Q0 training, rejecting component overlap, and rejecting in-sample or incomplete OOF assemblies. An independent artifact audit verifies all 40 teacher fit scopes' label and feature hashes, all 125 student artifacts, and the per-row seed assembly. Independently recomputed student NLLs agree with the reported values within 1.2×10^{-16}.

Results and limits. All four teachers have lower student NLL than ZERO, with nominal conditional intervals below zero (Table 22). Yet External 4D minus AudioBox has direct NLL difference -0.024740 and student difference $+0.001445$; 22D minus 10D has direct difference -0.016739 and student difference $+0.001721$. Both direct intervals exclude zero and both student intervals include it. The isolated follow-up therefore retains direct advantages without resolving a reverse teaching order or establishing equivalence. In the secondary alignment comparison, 22D has higher NLL than AudioBox by 0.001761 ($[0.000442, 0.003041]$); broad multi-task superiority is not supported.

The result does not retrospectively validate the original student or LF0/MERT analyses. Rebuilding selectors changes some budgets, and fresh tuning changes schedules; numerical environments also differ from the historical runs. A new-minus-old score difference would consequently not isolate label-reuse bias. T30's use of existing development votes and chosen configurations, its unadjusted conditional intervals, and its Q0-only coverage remain explicit limits. Original T7, official, T29 primary, and both reannotation outcomes are preserved.

Table 22: Fully fold-isolated Musicality comparisons (T30; 3,527 development votes). Entries subtract the second pipeline's negative log likelihood (NLL) from the first; negative favors the first. Q0 is the MuQ layer-fusion student; ZERO is matched human-only training; D counts teacher inputs. Nominal 95% intervals use 10,000 paired source-component resamples within outer fold, holding fits fixed; no multiplicity adjustment is applied. Dashes denote inapplicable direct contrasts: ZERO is a student control.

Contrast	Direct difference [interval]	Q0 difference [interval]
External 4D − AudioBox	−0.02474 [-0.03262, -0.01719]	+0.00145 [-0.00030, +0.00315]
22D − 10D	−0.01674 [-0.02349, -0.01010]	+0.00172 [-0.00028, +0.00370]
22D − AudioBox	−0.04686 [-0.05792, -0.03640]	+0.00170 [-0.00068, +0.00406]
AudioBox − ZERO	—	−0.00472 [-0.00812, -0.00128]
External 4D − ZERO	—	−0.00327 [-0.00634, -0.00025]
10D − ZERO	—	−0.00474 [-0.00808, -0.00140]
22D − ZERO	—	−0.00302 [-0.00569, -0.00042]

Table 23: Budgets and common schedules for fully fold-isolated student training (T30). Inner/outer pair counts are pseudo-supervision budgets, independently determined by the four-output teacher's confidence/agreement gate and capped at 11,631. LR/Decay mean learning rate/weight decay; mixed epochs use human and pseudo labels, followed by human-only epochs. Each fold's five training arms and five seeds share its selected schedule.

Fold	Inner pairs	Outer pairs	LR	Decay	Mixed epochs	Human epochs
0	11,631	11,631	0.0001	0	4	1
1	10,688	11,179	0.001	0	1	2
2	10,877	11,268	0.0003	0	2	4
3	11,009	11,613	0.001	0	1	2
4	11,631	11,631	0.001	0	1	1

D.5 SYMMETRIC TRAINING-SIDE CALIBRATION OF STUDENTS

Question and fixed scope. This post-hoc follow-up (T31 in the implementation index) asks whether the fold-isolated supervised students improve NLL over a ZERO ensemble given the same form of training-side scalar calibration. It reuses the existing 3,527 benchmark votes, 2,551 components, five outer folds, five seeds, and five arms from Appendix D.4. The scalar family, bounds, fitting scopes, and four-comparison testing family were frozen before the new inner fitting. No new human targets or official-test outcomes enter the experiment.

Inner fitting and temperature transfer. Using the notation of Equation 7, train all arms and seeds on F_k, using the existing inner teachers and pseudo supervision fitted only on F_k. Replay the common schedule previously selected on V_k, with that inner scope's matched budget, targets, and weights. Thus V_k is reused for schedule selection and temperature fitting; it is not an independent test. The new 125 inner student fits predict only V_k. Average the five seed probabilities before fitting one scalar for each arm a and fold k. With $z^{\mathrm{in}}_{k,a}(i) = \mathrm{logit}(\mathrm{clip}(\bar{p}^{\mathrm{in}}_{k,a}(i), 10^{-6}, 1-10^{-6}))$, fit positive inverse temperature $b_{k,a}$ by

$$b_{k,a} = \underset{b\in[0.05,20]}{\arg\min} \frac{1}{|V_k|} \sum_{i\in V_k} \left[\log(1+\exp(bz^{\mathrm{in}}_{k,a}(i))) - y_i bz^{\mathrm{in}}_{k,a}(i)\right], \qquad T_{k,a} = 1/b_{k,a}. \tag{8}$$

The objective is evaluated with a stable log-sum-exp implementation. There is no intercept. A root of the monotone derivative, or a boundary minimizer, solves this convex scalar problem; all-zero logits use $b = 1$. Apply the resulting temperature to the clipped logit of the corresponding frozen outer five-seed mean on H_k: $p^{\mathrm{cal}}_{k,a}(i) = \sigma(b_{k,a}\,\mathrm{logit}(\mathrm{clip}(\bar{p}^{\mathrm{out}}_{k,a}(i), 10^{-6}, 1-10^{-6})))$. The 125 outer fits are reused. No H_k label enters any fitting step, and predictions from other outer folds are never used to calibrate H_k.

The inner fits used CPU PyTorch with four threads per process; the outer fits used PPU-ZW810 accelerators. Model and training code, seeds, and schedules are preserved, but CPU and accelerator random streams need not yield identical fits. All arms share the new inner backend. This procedure transfers temperatures across inner/outer sample sizes and computational backends; it does not promise optimal calibration of the outer refits. The fitted temperatures range from 0.889 to 1.414, with no boundary solution.

Evaluation and results. All 125 new fits were completed before aggregate evaluation. The four primary comparisons for this follow-up are calibrated supervised-minus-calibrated-ZERO NLL. We use 10,000 paired component draws within outer fold, equal vote-row weights, and seed 20260915. Nominal percentile 95% intervals condition on predictions and fitted temperatures. Centered two-sided bootstrap p-values receive Holm adjustment across these four comparisons only. Raw-arm contrasts, each arm's calibration change, and three calibrated between-teacher contrasts are descriptive, separate from this family and all earlier primary tests.

Table 24 shows that AudioBox and 10D retain NLL gains over calibrated ZERO after this correction, approximately 0.77% and 0.84% of calibrated ZERO's NLL. The 4D interval includes zero; 22D has a negative nominal interval but does not pass the four-test correction. These outcomes do not establish equivalence for the latter two students or a ranking among the former two. All arms' NLL point

estimates decrease after calibration, but the conditional intervals for their own calibrated-minus-raw changes include zero. Brier point estimates increase slightly after calibration. Accordingly, these results concern NLL relative to the same calibrated baseline, not improvement of every proper loss from applying temperature scaling. All threshold decisions are unchanged.

Verification and limits. Eight targeted tests check scope selection, exclusion of held-out labels, swap symmetry, ensemble order, scalar fitting, component weighting, and multiplicity. The artifact audit checks 125 inner fit records and checkpoints, their teacher and schedule bindings, all inner and outer seed assemblies, and component separation. An independent bounded optimizer agrees with all 25 scalar fits in objective value within 1.2×10^{-16}. The standalone entry in Appendix D.6 refits the scalars and replays all reported statistical comparisons. The existing development population, chosen student and teacher configurations, temperature-transfer rule, and conditional intervals limit generalization. Retraining and calibration-fitting uncertainty, other calibration families, independent targets, and generation quality remain outside this experiment.

Table 24: Symmetric training-side temperature calibration of fold-isolated MuQ students (3,527 votes). ZERO is matched human-only training; D counts teacher inputs. Raw/Cal. denote uncalibrated/calibrated negative log likelihood (NLL). Negative contrasts favor supervised students over calibrated ZERO. Nominal 95% component-bootstrap intervals condition on fits and temperatures; p_{Holm} adjusts centered two-sided tests across the four comparisons. Dashes mark the reference row.

Teacher	Raw NLL	Cal. NLL	Cal. minus cal. ZERO [95% CI]	p_{Holm}
ZERO	0.487645	0.485866	—	—
AudioBox	0.482926	0.482138	-0.003728 [-0.006640, -0.000844]	0.038396
External 4D	0.484372	0.483183	-0.002683 [-0.005454, $+0.000064$]	0.056294
10D	0.482905	0.481774	-0.004092 [-0.007000, -0.001175]	0.022798
22D	0.484626	0.483143	-0.002723 [-0.005115, -0.000292]	0.054395

D.6 Applying and reproducing the framework

Reproduction index. Seven calculation entries cover the main reported comparisons. Their archives and distinct scopes are described below; each includes a README, dependencies, inputs, and reference results. These entries reproduce calculations rather than complete training pipelines.

Relation to evaluation and supervision methods. Music evaluators target quality, adherence and preference (Cideron et al., 2024; Liu et al., 2025; Grötschla et al., 2025). Our supervision comparisons connect distillation and weak supervision (Hinton et al., 2015; Sohn et al., 2020; Ratner et al., 2020) with teacher–learner alignment and student capacity (Sucholutsky et al., 2025; Gao et al., 2025). Work on visual generative models motivates examining representations for discriminative use (Li et al., 2023; Yang et al., 2026); our ACE-Step probe separately tests this possibility for music. These connections motivate the tested interfaces without supplying evidence for their music outcomes.

Reusable fitting contract. For a new evaluator, preserve the following interfaces independently of the model family. Supply processed condition inputs and an audio–condition scoring function for control; pairwise probabilities and grouped human targets for direct prediction; and a student trainer plus a teacher-to-supervision transform for teaching utility. Freeze intended comparisons, partitions, preprocessing, selectors, schedules or their training-side selection rules, and the uncertainty procedure before evaluation. The required flow is:

> Partition connected groups into outer training and evaluation scopes. Within each outer training scope, fit teacher transforms and pseudo supervision anew for every inner training scope used to select a schedule. Refit the selected pipeline on the outer training scope and predict only its excluded groups. Train ZERO with the same schedule, budget accounting, and ensemble rule. If calibrating students, fit the calibrator using only training-side validation predictions from models that exclude those validation labels from their training losses. Assemble held-out predictions by identity; compute control contrasts, direct contrasts, differences between students, and each student's change from ZERO as separate outputs. For feedback adaptation, form separate fitting and evaluation scopes for the new judgments. Reassess direct preference quality against frozen, matched-continuation, and training-side

calibration controls. A listener holdout also excludes test annotators from feedback fitting, selection, and calibration.

The calibration entry exposes reusable vector-based fitting, application, scope-checking, and component-resampling functions. The original entries replay this paper's numerical examples. A new scorer still requires its own inference and training adapters; recorded fitting scopes alone cannot validate an unseen implementation.

Prediction unit and computational scope. Teacher–student development predictions average the five seed models from the held-out student fold. Frozen T7, T29, and official assessments average all 25 fold–seed models. Feedback comparisons also average 25 members under the separate fitting scopes in Appendices D.14 and D.15. NLL is computed from the averaged probabilities, not averaged over member losses. In the Q0 supervision comparisons, teacher scoring and pseudo-target preparation occur offline. Supervision and ZERO share the same architecture, cached inputs, and ensemble size, so richer teachers do not themselves add student inference passes. Each Q0 prediction still requires its trained layer fusion, adapters, and heads.

For direct prediction, 22D adds SongEval and SongBench feature computation to 10D; 22 score outputs do not imply 22 model passes. The 17D ablation omits SongEval (Table 25), but its potential cost tradeoff is not independently evaluated. Recorded computation counts are not measurements of latency or cost effectiveness on matched hardware.

Shared reproduction conventions. For each of the six detailed statistical replay entries below, install its recorded dependencies and run `python3 reproduce.py --verify` in the extracted entry directory. Numerical references are read only after calculation and checked with absolute tolerance 10^{-12}. These entries use fixed model predictions; the calibration entry additionally refits temperatures. They do not rerun annotation QC, feature extraction, or model training. All seven archives exclude audio, prompts, original data identifiers, raw annotation returns, model checkpoints, and official-test rows. Their local availability and release conditions are stated at the end of this subsection.

Minimal reproduction entry. The locally prepared `ROLE_AUDIT_EXAMPLE_20260912.zip` contains `reproduce_role_audit/reproduce.py`, aggregate inputs, source hashes, and reference output. Running `python3 reproduce.py` with the standard library reconstructs strict/half-credit control BA, the common visible subset, T30 ensemble/seed contrasts, and T7 effect sizes. It also displays the archived official and T29 intervals. This entry reconstructs aggregate calculations; it does not recompute the displayed intervals.

Pair-level statistical reproduction. The separate `T29_STATISTICAL_REPRODUCTION_20260912.zip` contains a standalone NumPy replay for the original 999 scored T29 pairs and 483 source components. Its CSV contains mean human targets and seven frozen candidate probabilities; consecutive keys replace original pair and component identifiers. The two reannotation versions are covered by the three-version entry below.

The replay recomputes candidate NLL, the two primary contrasts, the seven prespecified Q0 contrasts, and the original reversal rule. It uses 50,000 whole-component draws with seeds 290835/290845, equal pair weights, centered two-sided tests, and seven-test Holm adjustment. Nominal percentile intervals remain distinct from adjusted p-values. The descriptive cross-role interval in Appendix D.12 comes from the original full analysis record; it is not among this entry's verified outputs.

T30 statistical replay and recorded fitting scopes. The locally prepared `T30_STATISTICAL_REPRODUCTION_20260914.zip` exports 3,527 benchmark vote-level targets, replacement row/component/fold keys, four direct teacher probabilities, and five seeds' Q0 probabilities for all five arms. With NumPy 1.26.0, the replay recomputes Musicality NLL, Brier and accuracy, the seven fixed student contrasts and three direct contrasts, using the original 10,000 within-fold component draws and seed 20260912. Each draw retains equal vote-row weights; intervals are nominal, conditional 95% intervals. Stored row and sorted-component orders are preserved.

The same entry checks recorded inner/outer component separation, fitting-label digests for 40 teacher scopes, the validation winner among each fold's six recorded schedule candidates, and common

schedule/budget bindings across all 125 student fits. The exporter checks all 125 prediction artifacts against the exported held-out-fold seed probabilities. Recorded memberships are exported from the artifacts, with identical ordered lists deduplicated. Original fitting and OOF-assembly source references, including training-flow excerpts, permit inspection of the code paths; they are not a standalone training implementation.

These checks validate recorded scope consistency separately from numerical reproduction. They do not replay historical execution or independently verify withheld features, pseudo selection, or model fitting. Alignment metrics are outside this entry.

Calibration fitting and statistical replay. `T31_CALIBRATION_REPRODUCTION_20260915.zip` exports benchmark vote targets, replacement row/component/fold keys, the five arms' outer ensemble means, and their inner-validation ensemble probabilities. With NumPy 1.26.0 and SciPy 1.11.3, the entry refits all 25 temperatures, applies them to outer predictions, and recomputes ten metric sets and 16 NLL contrasts, including the four-test Holm family. It uses the fixed draws, seed, vote weights, and conditional intervals specified in Appendix D.5. Recorded scopes for 125 inner fits, label digests, target bindings, and common schedules are also checked. It also supplies functions for fitting and applying the same calibrator to new aligned probability vectors.

Three-version statistical replay. `T29_THREE_ROUND_REPRODUCTION_20260916.zip` contains de-identified pair means, directional-majority targets, frozen probabilities for all 16 candidates, and consistent replacement pair/component keys for the original annotation and both reannotation versions. It reconstructs NLL, Brier score, accuracy, the direct/student contrasts, and the two student comparison families for each version, together with the 999-pair shared-set sensitivity. Whole-component bootstrap uses 50,000 draws with the recorded round-specific seeds; shared-set draws are paired across versions. The original and first-reannotation role estimates reproduce their archived values. Reannotation tests are post hoc, with Holm adjustment within each version and family, not across rounds.

The entry replays aggregate model comparisons; interrater reliability and construct correlations are outside its scope.

Human-feedback statistical replay. `T32_HUMAN_FEEDBACK_REPRODUCTION_20260916.zip` contains 2,397 de-identified pair means, nonmissing vote counts, directional-majority targets, fixed OOF probabilities for eight start–treatment combinations, and replacement pair, component, and fold keys. The program recomputes 24 NLL/Brier/accuracy cells and six NLL contrasts, including 20,000 fold-stratified paired component-bootstrap draws and the six-test Holm adjustment. It conditions on fitted models and temperatures; retraining uncertainty is outside its scope.

Listener-and-content holdout replay. `T33_PERSON_HOLDOUT_REPRODUCTION_20260916.zip` contains 473 de-identified single-listener targets, fixed predictions for the same eight start–treatment combinations, and replacement pair/component keys. It exports no listener identifiers. The replay reconstructs 24 NLL/Brier/accuracy values and all six NLL contrasts, including 20,000 paired component-bootstrap draws with seed 330916 and six-test Holm adjustment. The replay conditions on the reserved group and fitted predictions; it does not redo the person split, schedule selection, temperature fitting, or the per-listener and per-seed summaries.

All seven packages are collected in `ICLR2027_REPRODUCTION_SUPPLEMENT_20260922.zip`. Its manifest binds the archived 22 September 2026 manuscript snapshot and all seven calculation entries. Local preparation and any sharing follow the stated company-approval, scope, and rights conditions; local availability does not establish that the materials have been approved or supplied to reviewers.

D.7 Direct evaluator diagnostics

NLL is the proper-scoring endpoint. Adding either new scalar named *Musicality* does not reproduce the 22D reduction; the largest observed reduction occurs when the full multi-aspect heads are retained (Table 25).

Table 17 specifies the 10D base and its 22D expansion; Appendix D.1 gives the audio-only TuneJury configuration. All nested comparisons retain the same pair differencing, scaling, selection, and no-intercept calibration procedure. The inner selector chose both complete heads in all five outer folds.

Table 25: Nested out-of-fold feature-family ablation on 3,527 development votes (T14). Variants add SongEval/SongBench outputs to the same ten-input (10D) fusion. Dim. counts inputs; accuracy is in percent and negative log likelihood (NLL) is lower-is-better. Bold marks column leaders, not significance; "both complete heads" uses all outputs from both families.

Variant	Dim.	Accuracy	NLL
10D base	10	75.90	0.49881
+ SongEval Musicality	11	75.82	0.49877
+ all SongEval outputs	15	75.67	0.49691
+ SongBench Musicality	11	75.59	0.49694
+ all SongBench outputs	17	77.40	0.48471
+ both complete heads	22	**77.43**	**0.48201**

Table 26 compares simple aggregation alternatives. Across the 32,164 contents in the pseudo resource (Appendix D.8), the first five PCA components of the standardized outputs explain 88.90% of variance and the first ten explain 96.97%. The entropy-based covariance effective rank is $r_{\text{eff}} = \exp(-\sum_j v_j \log v_j) = 4.55$, where v_j is the fraction of variance explained by component j and $0 \log 0 = 0$. The 22 outputs thus provide correlated multi-view evidence. The family ablation above tests added outputs, not their statistical independence.

Table 26: Direct aggregation baselines on 3,527 development votes (T17). Accuracy is in percent; bold marks the highest point estimate. The best single output is selected within training folds. Equal RMS-scaled mean averages root-mean-square-scaled score differences; equal family mean first averages within evaluator families; sign majority votes over output directions. Learned fusions use 10 or 22 inputs and out-of-fold calibration.

Aggregator	Accuracy
Best single output	72.98
Equal RMS-scaled mean	73.09
Equal family mean	74.77
Sign majority	73.49
Learned 10D fusion	75.90
Learned 22D fusion	**77.43**

On Music Arena, 22D-minus-10D NLL is -0.002156 with user-cluster CI $[-0.017745, 0.014202]$. Thus the external result supports ranking accuracy, not a confirmed calibration or proper-score gain.

MuseCritic reports a Music Arena evaluation on 733 pairs (Zhuang et al., 2026), whereas ours uses 3,660 decisive comparisons. The archived comparability check could not recover the battle identifiers and A/B mapping needed to reconstruct a common population; we therefore do not rank these systems head to head.

D.8 Pseudo-supervision interface and schedules

T1 first fits the frozen no-intercept External-4D calibrator on all 3,527 development votes after its cross-validated gate passes. Its gold rule requires $\max(p, 1-p) \geq 0.90$ and agreement of all three nonzero raw family directions with the calibrated hard label. This selects 13,269 of 68,762 pairs: 11,631 in the inherited pseudo-training split and 1,638 in pseudo-validation. The full validation pool is excluded from the original teacher–student training; later feedback adaptation repartitions its annotated subset (Appendix D.14). AudioBox, 10D, and 22D selectors preserve the 11,631-pair training budget but rank their own calibrated confidences rather than reuse the T1 gold mask. These external teachers also use all development labels; the same pseudo target artifacts feed every student fold. This is the information path documented in Appendix D.3, not end-to-end OOF evaluation of the two-stage algorithm.

Unless stated otherwise, the selected-pair target is $\hat{y} = \mathbf{1}\{p \geq 0.5\}$ and the loss weight is $2|p - 0.5|$ without renormalization. Table 19 gives the G25 loss multiplier, mixed-batch schedule, human-only fine-tuning, and matched ZERO rule.

The first seed's original student-level inner search selected the schedules in Table 27; subsequent original Q0 comparisons reuse them without candidate-specific reselection. Each development row averages five seed predictions from its held-out student fold. Independent T7, official, and T29 pairs instead average all 25 seed-by-fold probabilities. LF0 and MERT replay the original Q0 schedules under separately frozen common-model protocols. T30 selects fresh fold-specific schedules and budgets (Table 23); it also averages only the held-out fold's five seeds. Student-fold exclusion in the original runs does not remove teacher-label reuse.

Table 27: Original schedules for the MuQ layer-fusion student (Q0), reused across historical teacher comparisons. G25 denotes a mixed phase with pseudo-loss multiplier 0.25, followed by human-only fine-tuning. Learning rate, weight decay, and both epoch counts were selected on inner validation. The fully fold-isolated follow-up uses the new schedules in Table 23.

Outer fold	Learning rate	Weight decay	Mixed epochs	Human-only epochs
0	10^{-4}	0	4	1
1	3×10^{-4}	0	2	1
2	10^{-3}	10^{-4}	1	3
3	10^{-3}	0	1	1
4	10^{-3}	0	1	1

D.9 Original student results and official-test protocol

Tables 28, 29, and 30 report the original core student/evaluator–interface combinations; Table 22 reports the fold-isolated comparisons, and Table 24 gives the calibrated comparisons. Utility is computed against the matched ZERO row for the same student. Bold marks the lowest NLL within each student block and is descriptive unless an accompanying interval is stated.

Table 28: Conditional student diagnostics on 3,527 development votes. Q0/LF0 use fused/final-layer MuQ features; MERT uses a MERT encoder; D counts teacher inputs. ZERO is matched human-only training, and Q0 self supplies its own pseudo targets. Accuracy is in percent; NLL is negative log likelihood; $10^3U = 1000(\mathrm{NLL}_{\mathrm{ZERO}} - \mathrm{NLL}_{\mathrm{student}})$, so positive is better. External teachers use all development labels, preventing end-to-end out-of-fold interpretation. Bold marks column leaders within each student using unrounded estimates, not significance.

Student	Teacher	Accuracy	NLL	10^3U
Q0	ZERO	76.84	0.48808	0.00
Q0	Q0 self	77.03	0.48643	1.65
Q0	External 4D	76.98	0.48370	4.38
Q0	AudioBox	**77.26**	**0.48234**	**5.74**
Q0	10D	77.06	0.48348	4.60
Q0	22D	77.01	0.48332	4.76
LF0	ZERO	76.47	0.49707	0.00
LF0	AudioBox	**76.95**	**0.49513**	**1.94**
LF0	22D	76.89	0.49514	1.94
MERT	ZERO	74.77	0.51957	0.00
MERT	AudioBox	74.45	0.51547	4.10
MERT	External 4D	74.74	0.51561	3.96
MERT	22D	**74.82**	**0.51502**	**4.55**

Official-test protocol. The protocol (T28) was frozen before official content or labels were accessed. It evaluated four direct scorers and 11 student ensembles on all 500 test pairs. Before inference, a schema amendment selected the ordered test rows from a 2,753-row JSONL by the unique `source=cmi-arena-annotation` field. One empty prompt maps to a fixed 512D

Table 29: Frozen MuQ layer-fusion students (Q0) on the independent human student test (T7). Negative log likelihood (NLL) uses 1,195 pair-mean targets; accuracy (%) uses 1,134 three-vote majorities. Utility is $1000(\mathrm{NLL}_{\mathrm{ZERO}} - \mathrm{NLL}_{\mathrm{student}})$, relative to matched human-only training. HARD/SOFT use binary/probabilistic four-output teacher targets; MT pool/Q0 label uses that teacher's selected pairs with self targets. Predictions average 25 models. Bold marks column leaders, not significance.

Teacher/interface	Accuracy	NLL	10^3U
ZERO exact	**85.10**	0.40985	0.00
Q0 self	84.48	0.41055	−0.69
External 4D/HARD	84.39	0.40721	2.64
AudioBox	84.39	**0.40514**	**4.71**
MT pool/Q0 label	84.30	0.40809	1.76
External 4D/SOFT	84.39	0.40684	3.01

zero vector; 29 overlength texts use native right truncation. A reporting-only amendment added predeclared contrasts without changing predictions, metrics, or decisions. Official outcomes were not used for fitting, selection, calibration, thresholding, or filtering.

The predeclared joint ordering test required both higher External-4D direct accuracy than AudioBox and lower AudioBox-supervised Q0 NLL than External-4D supervised Q0 NLL. Both 95% intervals had to exclude zero in the specified directions. Unlike the earlier comparison of development accuracy with T7 student loss, this test uses one evaluation population. The separate common-target human test in Appendix D.12 instead uses NLL for both roles.

Table 30: Frozen students on the 500-pair official test: Musicality preferences and secondary condition Alignment. Q0/LF0 use fused/final-layer MuQ features; MERT uses a MERT encoder; D counts teacher inputs. ZERO is matched human-only training. Accuracy is in percent; negative log likelihood (NLL) is lower-is-better. Each prediction averages 25 models. Predeclared Musicality gain and teacher-order intervals include zero; bold marks point-estimate leaders within each student/task, not significance.

		Musicality		Alignment	
Student	Teacher	Accuracy	NLL	Accuracy	NLL
Q0	ZERO	80.60	0.45457	**74.40**	0.529154
Q0	AudioBox	80.60	0.45220	73.80	0.530037
Q0	External 4D	80.60	0.45227	73.80	0.529007
Q0	22D	**80.80**	**0.45150**	**74.40**	**0.528434**
LF0	ZERO	77.80	0.47650	**74.60**	**0.520959**
LF0	AudioBox	77.80	0.47612	74.00	0.524381
LF0	22D	**78.60**	**0.47526**	73.40	0.523830
MERT	ZERO	**77.20**	0.50410	**72.20**	**0.548907**
MERT	AudioBox	77.00	0.50333	71.00	0.552176
MERT	External 4D	76.40	**0.50326**	71.20	0.552583
MERT	22D	76.60	0.50397	71.60	0.552564

The official pipeline-minus-ZERO NLL estimates are -0.002366 for Q0/AudioBox, -0.000381 for LF0/AudioBox, -0.001237 for LF0/22D, -0.000771 for MERT/AudioBox, -0.000840 for MERT/4D, and -0.000128 for MERT/22D. Their connected-component 95% intervals all include zero. The Q0 22D-minus-AudioBox estimate is -0.000702 (CI $[-0.004398, 0.002971]$).

No prospective minimum-detectable effect was recorded, so retrospective power is omitted. The Q0 4D-minus-AudioBox co-primary and 22D-minus-AudioBox secondary intervals span 0.005409 and 0.007369 NLL, respectively, admitting effects comparable to development/T7 contrasts. The other co-primary endpoint is direct 4D-minus-AudioBox accuracy, whose interval is $[-3.018, 3.024]$ percentage points. "One-shot" limits adaptation, not interval width.

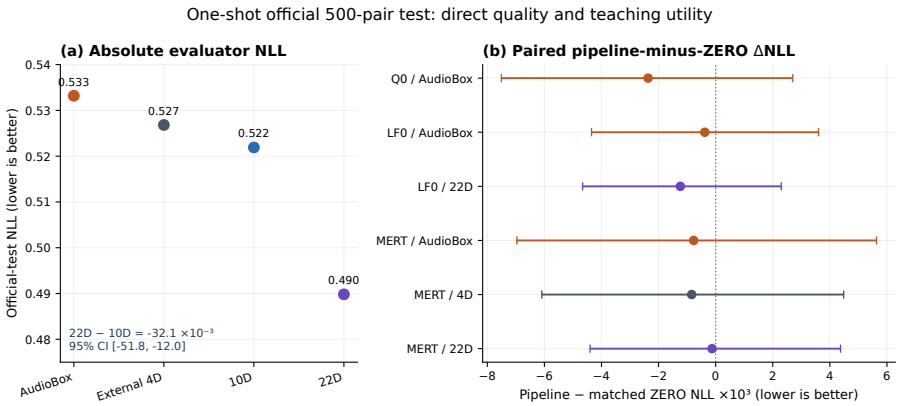

Figure 4: Frozen Musicality predictions on the official 500-pair test. (a) Direct negative log likelihood (NLL; lower is better), with the prespecified 22D-minus-10D contrast; D counts fused inputs. (b) Six prespecified NLL contrasts between supervised students and their human-only (ZERO) controls. Q0/LF0 use fused/final-layer MuQ features; MERT uses a MERT encoder. Labels give student/teacher; predictions average 25 models. All contrasts and confidence limits use 10^{-3} NLL units. In (b), negative favors supervision; its nominal 95% component-bootstrap intervals include zero without establishing equivalence.

D.10 Quality–utility map and student dependence

Figure 5 plots the original conditional development comparisons. Across five Q0 selectors, the descriptive Spearman correlation between evaluator accuracy and negative student NLL is -0.30, with six of ten order disagreements. LF0 22D-minus-AudioBox NLL is $+0.000005$ (CI $[-0.001122, 0.001105]$), and the LF0-minus-Q0 interaction is -0.000974 (CI $[-0.003368, 0.001425]$). MERT 4D-minus-AudioBox is $+0.000137$ (CI $[-0.000993, 0.001286]$). These intervals do not identify a population-level interaction or cross-student teacher–interface order.

D.11 T7 human evidence and execution audit

T7 was frozen before sampling or annotation. Its 1,200 pairs cover all 739 pseudo-validation components, with at most two pairs per component and no pseudo-training audio overlap; selection was blind to model scores, development labels, and legacy preferences. The blinded, independently reoriented interface asked which audio was more musically pleasing and coherent and allowed cannot judge. Each of three logical slots received six 200-pair packets plus 20 hidden repeats per packet: 3,600 primary and 360 repeat judgments. Every pair retained three distinct assignments.

Nine contributors staffed the intended three logical slots, two packets each; this deviation was documented before candidate unblinding and did not change the item-level design. Logical-slot hidden-repeat consistency was 88.33%, 86.67%, and 84.17%; individual consistency ranged from 80% to 95%. All completion and cannot-judge gates passed. Five pairs had no valid vote; 1,195 had at least one valid vote and entered the primary pair-averaged NLL analysis, totaling 3,513 valid votes. Of these, 1,134 pairs had all three valid votes and a binary majority and enter the accuracy analysis. NLL uses all 1,195 scored pairs. For each retained pair, the target is the mean of its valid binary votes. NLL averages the resulting losses equally over pairs, so a pair with fewer valid votes receives the same weight as one with three.

HARD-minus-ZERO NLL is -0.00264 (95% component-bootstrap CI $[-0.00490, -0.00039]$), and HARD-minus-Q0-self NLL is -0.00334 ($[-0.00516, -0.00148]$), although HARD's majority accuracy is lower. HARD-minus-AudioBox NLL is $+0.00207$ ($[+0.00060, +0.00357]$). T7 therefore supports external pseudo supervision under this interface, without establishing a multi-teacher-specific attribution.

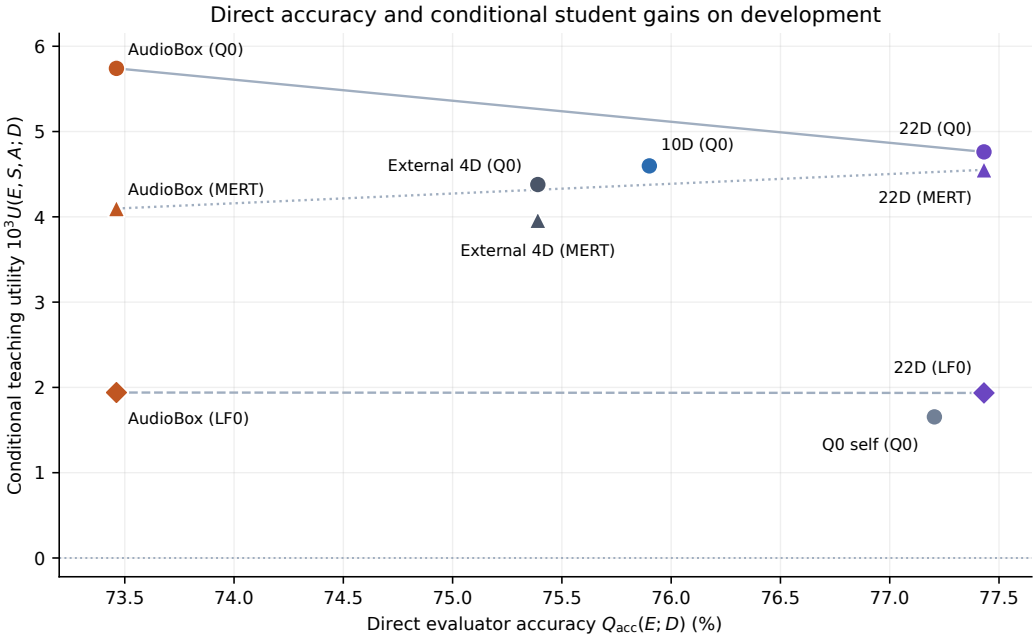

Figure 5: Direct accuracy and conditional student gains on 3,527 development votes. Horizontal position is direct prediction accuracy from out-of-fold evaluation; vertical position is $1000(\mathrm{NLL}_{\text{human-only}} - \mathrm{NLL}_{\text{supervised}})$, so higher means lower student log loss. Circles/diamonds use fused/final-layer MuQ students (Q0/LF0); triangles use MERT. D counts teacher inputs. Lines connect AudioBox and 22D within a student, not fitted trends. For pseudo supervision, external teachers use all development labels, while Q0 self-targets are fold-specific. The student gains are therefore conditional diagnostics, distinct from the fully fold-isolated results in Table 22.

D.12 T29 ORIGINAL SAME-PAIR HUMAN AND CONSTRUCT AUDIT

T29 was frozen after a label- and score-blind feasibility audit. It selects 2,400 pseudo-validation comparisons from 570 source components, excludes every exact T7 pair and all development and official audio, and assigns every pair to four of 24 distinct people, confirmed by the project owner. Each person receives two 200-pair packets with 20 independently reoriented repeats per packet. Direct and student predictions were frozen without reading returned responses; each student probability is the arithmetic mean of the same 25 frozen fold–seed models used in earlier stages.

QC on the original returns retains 14 worker slots. Ten fail at least one substantive worker gate: eight fail the 80% hidden-repeat consistency threshold, four exceed the 8% cannot-judge ceiling, and three supply fewer than 30 valid repeat pairs; these counts overlap. All 48 files have exact schemas and complete packet rows. An initial exact duration comparison also marked 680 rows short by at most 20.334 microseconds. Before any human–prediction join, the audit recorded the original QC hash and added a fixed 0.1-ms numerical tolerance for the browser-decoder/container boundary. Under exact comparison 7/24 slots pass all gates, versus 14/24 with this tolerance. The UI enforced its browser-side listening gate, but the artifact cannot independently reconstruct that internal endpoint. Required listening durations, the 95% worker fraction, and all non-listening QC thresholds remain unchanged.

After excluding all rows from failed slots, 1,000 pairs have at least three passing assignments. One has only cannot-judge responses, leaving 999 pairs and 3,341 nonmissing votes over 483 source components. Nominal Krippendorff α is 0.578, raw pairwise exact agreement is 76.05%, and the tie-vote rate is 6.41%. Fourteen slots pass QC, but only 13 contribute a nonmissing vote to retained packets because both packets assigned to one passing slot fall below the pair-level three-passer rule. Analyses track participants with anonymous slot IDs.

For T29 and both reannotation versions, canonical audio-A, audio-B, and tie votes map to 1, 0, and 0.5; cannot-judge votes are missing. Each target is the mean of the remaining votes, and fractional NLL gives each scored pair equal weight. The three-passer gate counts complete assignments, including documented cannot-judge responses; it does not guarantee three nonmissing votes. Accuracy instead uses the majority of directional votes, excluding ties and unresolved directional splits. This differs from T7 accuracy, which requires three valid binary votes. Exact model-probability ties receive half accuracy credit in all three versions. Their NLL probabilities are clipped to $[10^{-7}, 1-10^{-7}]$, as in T7. T30 and its T31 calibration follow-up use $[10^{-6}, 1-10^{-6}]$ for NLL evaluation (Appendices D.4 and D.5). Numerical NLL applies Equation 3 to these clipped probabilities; the theory concerns the unclipped log loss.

Descriptive reliability estimation. Nominal Krippendorff α uses coincidence weights $1/(m_i - 1)$ for items with $m_i \geq 2$ nonmissing votes and pairable-value marginals, following Krippendorff (2011). This normalization accounts for varying numbers of raters; single-vote items do not enter the reliability calculation. Raw pairwise exact agreement is reported separately. Reliability is descriptive and does not change the QC gates, preference targets, or model comparisons.

Table 31 reports the original direct and student metrics on these targets.

The primary direct NLL difference (External 4D minus AudioBox) is -0.026970, with a 95% source-component interval of $[-0.035546, -0.018411]$. The matched Q0 NLL difference is $+0.000144$, with interval $[-0.001440, 0.001662]$. Their prespecified intersection–union test of an opposite Q0 ordering is therefore not satisfied on the teaching condition. Directional-vote-only and cross-classified component–worker sensitivities preserve the direct interval and retain a teaching interval crossing zero. In the Q0 Holm family, the 10D-minus-AudioBox NLL delta is -0.001672 ($p_{\mathrm{Holm}} = 0.0318$), while 22D has higher NLL than 10D by $+0.002392$ ($p_{\mathrm{Holm}} = 0.0228$); other family tests do not reject. The three LF0/MERT teacher–interface-order intervals also cross zero. Pooling T7 and T29 only as prespecified secondary evidence gives 4D-minus-AudioBox $+0.001195$ ($[0.000093, 0.002297]$); 483 shared components preclude treating this as a new component-independent replication.

Descriptive comparison of direct and student contrasts. Let Δ_{direct} and $\Delta_{\mathrm{student}}$ denote the External-4D-minus-AudioBox NLL contrasts for direct evaluators and their Q0 students, respectively, on the same 999 human targets. The stored paired analysis subtracts these contrasts within each source-component bootstrap draw. It gives $\Delta_{\mathrm{student}} - \Delta_{\mathrm{direct}} = +0.027114$, with a nominal 95%

Table 31: Frozen direct scorers and MuQ layer-fusion students (Q0) on 999 common human pair-mean targets from the original annotation (T29). NLL is negative log likelihood (lower is better); accuracy uses 947 resolved directional majorities. Student predictions average 25 models; candidate names identify their teachers, D counts teacher inputs, and ZERO is matched human-only training. These judgments were unused in fitting the listed models.

Role	Candidate	NLL	Accuracy (%)
Direct	AudioBox	0.45324	82.47
Direct	External 4D	0.42627	83.63
Direct	10D	0.41545	85.01
Direct	22D	0.39425	86.91
Q0 student	ZERO	0.38598	88.38
Q0 student	AudioBox	0.38533	87.65
Q0 student	External 4D	0.38548	87.75
Q0 student	10D	0.38366	87.96
Q0 student	22D	0.38605	88.07

interval $[+0.018849, +0.035434]$. This descriptive comparison identifies a difference between the signed NLL contrasts for these frozen prediction systems, without serving as an additional success gate. The predeclared reversal rule still requires both primary directional intervals to pass, which they do not. The result establishes neither the opposite student ranking nor equivalence between teachers, and does not identify the cause of the difference.

Post-hoc QC-selection diagnostics. Because passing slots are packet-bound, the primary gate retains five whole packets and excludes seven whole packets rather than selecting isolated pairs. Table 32 therefore treats the frozen primary result as the only confirmatory row and the alternatives as stress tests. Lowering only the pair gate to at least two passing slots adds four packets; the unweighted all-slot row includes data from failed slots and is not an inferential replacement.

Table 32: Original common-target evaluation (T29): sensitivity to quality-control (QC) selection. Thresholds count passing annotators assigned to each pair; scored pairs also require a nonmissing vote. Comp. counts shared-source components. Differences are External 4D minus AudioBox negative log likelihood (NLL), directly and through their MuQ students (Q0); negative favors External 4D. Brackets are nominal 95% source-component bootstrap intervals. Only the three-passer row is prospective; all-slot data include failed annotators.

Cohort	Eligible / evaluated	Comp.	Direct Δ [95% CI]	Q0 Δ [95% CI]
Frozen ≥ 3	1,000 / 999	483	$-0.026970\,[-0.035546, -0.018411]$	$+0.000144\,[-0.001440, +0.001662]$
Post-hoc ≥ 2	1,800 / 1,795	543	$-0.030960\,[-0.038512, -0.023585]$	$+0.000437\,[-0.000771, +0.001620]$
All slots, unweighted	2,400 / 2,398	570	$-0.027434\,[-0.033706, -0.021234]$	$+0.000588\,[-0.000400, +0.001549]$

The packet groups by passing-slot count are: one—01, 04, 08; two—02, 06, 09, 12; three—05, 07, 10; and four—03, 11. Thus only the last two groups enter the frozen primary cohort. Comparing its 1,000 QC-eligible pairs with the 1,400 excluded pairs, the largest absolute standardized mean difference is 0.086 across pair duration summaries, full-selection component multiplicity, four frozen direct confidence magnitudes, and frozen 4D–AudioBox disagreement. Task-type total variation is 0; generator-endpoint total variation is 0.0376, using label-free provenance that maps every selected endpoint. The groups contain 483 and 524 components, of which 437 overlap.

On the fixed primary pair set, leave-one-passing-slot point estimates range from -0.029462 to -0.025221 directly and from -0.000314 to $+0.000479$ for Q0. Leaving out one of the five retained packets gives ranges $[-0.029676, -0.024238]$ and $[-0.000476, +0.000726]$, respectively. These diagnostics preserve a negative direct delta and an unresolved teaching order, but neither prove retained/excluded exchangeability nor remove packet-selection uncertainty. No quality-weighted all-worker result is added because no continuous weighting rule was frozen; defining one after outcome access would introduce analyst discretion.

The embedded subset has 200 soft targets for each additional aspect. Human agreement is moderate: $\alpha = 0.607$ for melody/structure, 0.582 for arrangement/instrumentation, and 0.487 for audio quality.

Table 33 tests the seven mappings fixed before annotation. Holm correction applies to the two-sided component-bootstrap directional-accuracy tests; correlations are descriptive.

Table 33: Original model-blind human-aspect audit (T29): seven predeclared mappings from human aspects to evaluator outputs. Accuracy uses n_{acc} resolved directional majorities; Spearman rank correlation ρ compares output score differences with all 200 pair-mean human targets per aspect. Higher values indicate alignment with the elicited judgment, not separation from other aspects.

Human aspect	Frozen output	n_{acc}	Acc. (%)	ρ
Melody/structure	SongBench Melody	192	82.29	0.756
Melody/structure	SongBench Structure	192	83.85	0.786
Arrangement/instrumentation	SongBench Arrangement	193	82.90	0.775
Arrangement/instrumentation	SongBench Instrumental	193	79.27	0.726
Audio quality	SongEval Clarity	187	79.68	0.681
Audio quality	SongEval Naturalness	187	77.54	0.630
Audio quality	SongBench Mixing	187	77.54	0.745

All seven adjusted accuracy p-values are at most 0.00014. This supports broad alignment with the three elicited human judgments on the selected subset. It does not identify causal feature effects, prove independence among 22 outputs, or show that any one output improves a student.

In a post-hoc discriminant diagnostic, the three human aspect targets are strongly associated with the human overall-musicality target (Spearman $\rho = 0.963$, 0.941, and 0.920), so high raw matched correlations alone do not establish specificity.

The mean raw correlation is 0.728 for the seven prespecified matches and 0.731 for the 14 nonmatches; the corresponding partial means are 0.185 and 0.183. A prespecified aspect is the largest partial correlate for three of seven outputs. These descriptive results offer no clear discriminant-separation evidence. The supported construct claim therefore remains alignment with three broad elicited judgments, not validation of seven distinct psychological constructs.

Table 35 compares the original and both revised annotation versions; Table 34 summarizes retained people, pairs, and votes across all three versions.

D.13 Human-target stability across two reannotation rounds

Version policy and coverage. We compare the original annotation with both subsequent reannotation rounds on the same selected pool. The first reannotation is recorded as R2 and the second as R3. R3 returns complete assignments for the three remaining failed slots. Each designated slot uses its entire latest return; the other 21 retain their R2 records. Old and revised answers are never pooled as additional votes, and hidden repeats remain QC items only. All numerical gates, candidate models, calibrators, and predictions stay fixed. The response-only R3 QC decision was sealed before the new prediction join.

In the second reannotation, slots 06 and 14 pass every gate, with repeat agreement $33/37 = 89.19\%$ and $33/38 = 86.84\%$ and cannot-judge fractions 5.25% and 3.25%. Slot 12 remains below the 80% repeat threshold ($24/40 = 60.00\%$); all its judgments are excluded. The final consolidated version therefore retains 23 of the same 24 people. All 2,400 selected pairs meet the three-passing-worker completion gate; three have no nonmissing overall preference, leaving 2,397 targets, 9,020 nonmissing votes, and 570 source components. These are overlapping data versions, not additive samples or independent replications. The original 999-pair prospective analysis remains unchanged; both reannotation rounds are post-hoc sensitivity evidence.

Table 34: Coverage of the same preference-annotation pool across three versions. The people column counts quality-control (QC) passers out of 24; not every passer contributes to retained pairs. Eligible pairs have at least three complete assignments from passers; scored pairs also have a nonmissing overall preference. Components group shared sources; vote counts include nonmissing judgments on scored pairs. Revised answers replace earlier answers, and rows must not be summed.

Version	People	Eligible / scored	Components	Votes
Original	14/24	1,000 / 999	483	3,341
First reannotation	21/24	2,200 / 2,197	563	7,854
Second reannotation	23/24	2,400 / 2,397	570	9,020

Table 35: Direct and MuQ-student (Q0) contrasts across annotation versions. Differences are four-output External 4D minus AudioBox negative log likelihood (NLL); negative favors External 4D or its student. Brackets are nominal 95% intervals from 50,000 source-component bootstrap draws. Each row uses the same human pair-mean targets for both roles and unchanged model predictions. Rows overlap; only the original is the frozen prospective comparison.

Version	Pairs	Direct Δ [95% CI]	Q0 Δ [95% CI]
		-0.026970	$+0.000144$
Original	999	$[-0.035546, -0.018411]$	$[-0.001440, +0.001662]$
		-0.029540	$+0.000392$
First reannotation	2,197	$[-0.036269, -0.022869]$	$[-0.000712, +0.001489]$
		-0.031155	$+0.000458$
Second reannotation	2,397	$[-0.037747, -0.024654]$	$[-0.000616, +0.001509]$

Both reannotation versions retain External 4D's direct advantage, while the matched student intervals admit either order. Enlarged coverage therefore preserves the distinction between the measured direct advantage and the unresolved student comparison. It does not establish equivalence or satisfy the original reversal criterion.

Holding pair coverage fixed. All three scored sets share 999 pairs from 483 source components. Table 36 holds those pair identities and predictions fixed while reconstructing targets from each version's passing votes. This separates changes in target composition from expanded pair coverage; it does not isolate a causal effect of re-listening. The same pattern persists: a direct advantage and a student interval spanning zero. A common bootstrap seed pairs draws across these three sensitivity rows; the original frozen intervals remain those in Table 35.

Table 36: Annotation sensitivity on the same 999 pairs and 483 source components in every row. Differences are four-output External 4D minus AudioBox negative log likelihood (NLL), directly and through their MuQ students (Q0); negative favors External 4D. Only quality-control-retained human pair-mean targets change; predictions stay frozen. Nominal 95% intervals use 50,000 shared component-bootstrap draws. Rows are dependent, post-hoc comparisons.

Target version	Direct Δ [95% CI]	Q0 Δ [95% CI]
	-0.026970	$+0.000144$
Original	$[-0.035665, -0.018252]$	$[-0.001434, +0.001659]$
	-0.025784	$+0.000333$
First reannotation	$[-0.034354, -0.017212]$	$[-0.001220, +0.001814]$
	-0.025922	$+0.000480$
Second reannotation	$[-0.034467, -0.017424]$	$[-0.001100, +0.001974]$

Candidate-specific effects. Changing annotation versions changes both the target mixture and the evaluation population; comparisons therefore use identical targets within each version rather than losses across versions. On the final version, the 10D Q0 student has lower NLL than AudioBox, ZERO, and 22D under the seven-comparison Holm family (Table 37). The 22D–AudioBox student difference does not pass that correction in any version. The 10D–ZERO comparison passes in both reannotation versions but not the original cohort. These results delimit candidate-specific evidence;

they do not imply a stable total ordering across data versions or backbones. All reannotation-family tests remain post hoc; adjustment is within each stated family and does not correct repeated inspection across rounds.

Table 37: Secondary teacher comparisons within each student. Latest differences are first-minus-second negative log likelihood (NLL) on 2,397 second-reannotation targets; negative favors the first. Brackets are nominal 95% source-component bootstrap intervals. All three p columns are Holm-adjusted within version: seven MuQ layer-fusion (Q0) tests, or separately three final-layer MuQ (LF0)/MERT tests. D counts teacher inputs; ZERO is human-only training. Reannotation comparisons remain exploratory.

Comparison	Latest Δ [95% CI]	Original p	First p	Second p
Q0: 22D − AudioBox	$+0.001159$ $[-0.000309, +0.002613]$	1.0000	0.2496	0.3295
Q0: 10D − AudioBox	-0.001660 $[-0.002462, -0.000875]$	0.0318	0.0010	0.0004
Q0: 22D − 10D	$+0.002819$ $[+0.001695, +0.003943]$	0.0228	0.0001	0.0001
Q0: AudioBox − ZERO	-0.001716 $[-0.003667, +0.000191]$	1.0000	0.2365	0.3206
Q0: 4D − ZERO	-0.001258 $[-0.002819, +0.000267]$	1.0000	0.2365	0.3295
Q0: 10D − ZERO	-0.003376 $[-0.005179, -0.001603]$	0.3807	0.0014	0.0014
Q0: 22D − ZERO	-0.000557 $[-0.001848, +0.000717]$	1.0000	0.2683	0.3954
LF0: 22D − AudioBox	$+0.001480$ $[+0.000718, +0.002236]$	0.4183	0.0005	0.0005
MERT: 4D − AudioBox	$+0.000043$ $[-0.000753, +0.000820]$	0.4292	1.0000	1.0000
MERT: 22D − AudioBox	$+0.000322$ $[-0.000760, +0.001398]$	0.4292	1.0000	1.0000

Reliability and construct specificity. The final version has 76.38% raw pairwise exact agreement and a 6.14% tie-vote rate. Nominal reliability remains similar across versions (Table 38); passing within-person repeat checks does not eliminate disagreement between listeners. The final version supplies 479 scored pairs for the three additional aspects. Mean matched/nonmatched raw correlations are 0.704/0.704, and partial correlations controlling overall musicality are 0.163/0.147. The prespecified aspect has the largest partial correlation for four of seven outputs. These descriptives support broad alignment without establishing distinct constructs or explaining the student effects.

Table 38: Human-label reliability across overlapping annotation versions. Cells give scored pairs / nominal Krippendorff α (1: perfect agreement; 0: chance-level agreement) for A/B/tie categories. Coincidence normalization handles varying nonmissing-rater counts; single-vote pairs are excluded from α but can enter scored-pair counts. The versions are not independent samples.

Human target	Original	First reannotation	Second reannotation
Overall musicality	999 / 0.578	2,197 / 0.568	2,397 / 0.567
Melody/structure	200 / 0.607	439 / 0.555	479 / 0.544
Arrangement/instrumentation	200 / 0.582	439 / 0.514	479 / 0.511
Audio quality	200 / 0.487	439 / 0.487	479 / 0.479

Scope and reproduction. The same frozen predictions underlie every version, so these comparisons measure sensitivity to human-target revision and coverage. Feedback adaptation uses the final targets under separate fitting and evaluation scopes (Appendix D.14). The reliability calculation, full candidate metrics, shared-pair estimates, construct matrices, and leave-one-worker/packet influence diagnostics are recorded in `EVALUATION_THREE_ROUNDS_20260916.json`. `analyze_t29_three_rounds.py` verifies the original primary estimates and computes reliability consistently across versions. A de-identified statistical export supports offline replay of the three-round NLL, Brier, accuracy, role contrasts, and student comparison families (Appendix D.6).

D.14 HUMAN-FEEDBACK ADAPTATION: DESIGN AND COMPLETE RESULTS

Human-feedback adaptation, indexed as T32 in the execution archive, tests whether the collected human judgments improve a music preference scorer when used as supervision. It uses standard pairwise preference learning, not a new training algorithm. Its endpoint is the adapted scorer's direct preference quality Q; neither its downstream teaching utility U nor a generation policy is evaluated. The design was fixed before adaptation, but the cohort outcomes and model comparisons had already been inspected. The choice of ZERO and 10D starts was therefore informed by prior results, making this a post-hoc exploratory experiment.

Targets and data reuse. We use the final consolidated annotation version, formed by replacing each designated slot's complete return and then applying the unchanged QC gates. This yields 9,020 nonmissing overall-musicality votes from 23 people on 2,397 pairs in 570 source components, exactly the final aggregate targets in Appendix D.13. No person–task vote is counted twice; hidden repeats remain QC-only. A vote for A is 1, a vote for B is 0, a tie is 0.5, and an unjudgeable response is missing. Three pairs with no nonmissing vote are excluded. For each pair i, the target q_i is the mean of its nonmissing votes. New pairs have equal loss weight regardless of their number of votes; votes are averaged within a pair. The tie convention supplies fractional supervision, not a separate tie-probability model. Aspect-specific judgments are not used in this adaptation. This is the primary feedback-adaptation analysis. It holds out content while retaining the same annotator pool across fitting and evaluation. The supplementary analysis in Appendix D.15 also excludes test annotators from the new feedback used for fitting and selection.

Content isolation. The pairs are drawn from the existing pseudo-validation pool. Component membership is disjoint from pseudo-training, and audio hashes are disjoint from the original human development data. Five new outer folds each contain 114 source components; their test-pair counts are 480, 480, 479, 479, and 479. Splitting uses only identities and component sizes, with seed 20260916. No source component, audio endpoint, or audio hash crosses a fold boundary. Within each outer training scope, 365 components form the inner fit set and 91 form inner validation; the latter contains 384 pairs. Each final prediction uses only models and temperatures fitted without its outer fold's labels. This separation tests held-out content from the same annotator pool, not generalization to unseen people. The overlapping independent student-test population (T7) is not reused as a fresh test. The official benchmark is neither read nor reevaluated. Historical frozen-model estimates retain their original predictions and targets; the new cross-validation estimates are reported separately.

Model and objective. For each start, we reuse all 25 original checkpoints: five development folds and five seeds (260114, 260214, 260314, 260414, and 260514). ZERO used human labels without pseudo loss; the 10D start additionally received ten-output teacher supervision. The original human replay set for each checkpoint contains only that checkpoint's original outer-training Musicality rows. The original held-out development labels are not added to its replay set. The trainable score $s_\theta(x)$ applies the original rank-64 layer fusion, audio adapter, and Musicality head to cached frozen MuQ features. These modules contain 1,066,689 trainable parameters. MuQ, its MuLan projection, other heads, and condition adapters are frozen. Updating a shared representation can still change other task outputs, so preservation of alignment performance is untested. For pair i with audio inputs (x_{a_i}, x_{b_i}),

$$\begin{aligned} p_{\theta,i} &= \sigma(s_\theta(x_{a_i}) - s_\theta(x_{b_i}))\,, \\ \ell_i(\theta) &= -q_i \log p_{\theta,i} - (1-q_i)\log(1-p_{\theta,i}), \\ \mathcal{L}_{\text{feedback}}(\theta) &= (1-\alpha)\mathbb{E}_{i\sim D_{\text{old}}}[\ell_i(\theta)] + \alpha\mathbb{E}_{i\sim D_{\text{new,train}}}[\ell_i(\theta)]. \end{aligned} \tag{9}$$

On old-data rows, q_i is the original binary vote; on new pairs, it is the mean target defined above. The old expectation is uniform over vote rows and the new expectation over pairs. Binary cross-entropy is affine in its target, so using q_i exactly equals averaging the individual fractional losses within that pair. The objective retains disagreement rather than replacing it with a hard majority.

Matched controls. Four treatments start from the same checkpoints. *Frozen* retains their predictions. *Old data only* continues training on two old-data minibatches per step. *New feedback* replaces the second minibatch with new outer-training pairs, using Equation 9. Both training arms use two 64-row minibatches, the same α, optimizer, and step budget. *Temperature only* fits one positive scalar per start and new outer fold to the frozen ensemble's new outer-training probabilities. The inverse temperature

is bounded to $[0.05, 20]$ and there is no intercept. All arms average probabilities over the same 25 checkpoints or their adapted descendants. Positive temperature scaling preserves threshold decisions; it is fitted to the ensemble probability, not separately to its members.

Training-side selection. For each start and new outer fold, the fixed original fold-0, seed-260114 checkpoint represents schedule selection. Six inner feedback fits cross learning rates $\{10^{-5}, 3 \times 10^{-5}, 10^{-4}\}$ with $\alpha \in \{0.25, 0.5\}$. The smallest inner-validation NLL across epochs 0–12 selects the schedule; ties prefer fewer epochs, then lower learning rate, then lower α. Both training arms and all 25 initializations share that schedule. Old data only therefore uses no new-label gradients, but its schedule indirectly uses new-label validation through the feedback arm. This estimates an effect under a common schedule, not the difference between independently optimized treatments. Final training restarts from each original checkpoint on the entire new outer-training scope, using AdamW, weight decay 10^{-4}, gradient-norm clipping at 1, no learning-rate scheduler, and $\lceil n_{\text{new,train}}/64 \rceil$ steps per epoch in both arms. For new fold k and original checkpoint fold j (both indexed 0–4), the training seed is $s_{\text{base}} + 3{,}200{,}000 + 10{,}000k + 100j$, where s_{base} is one of the five source seed bases above. The experiment completes 60 inner trials, 500 final adaptation fits, and ten temperature fits. Every final fit updates the designated trainable parameters.

Estimation and results. Each pair contributes exactly one outer-held-out probability per treatment. Primary NLL equally weights the 2,397 pair means. The main contrasts compare new feedback with old-data continuation for each start; comparisons with frozen and temperature-only predictions complete one planned family of six NLL tests. We use 20,000 paired component-bootstrap draws, stratified by new outer fold, with seed 320916. Percentile 95% intervals are nominal and conditional on the fitted cross-validation predictions. Centered two-sided bootstrap tests use a finite-draw correction, followed by Holm adjustment over all six contrasts. These intervals exclude uncertainty from repeating training and do not remove the post-hoc choice of this follow-up. Final evaluation clips probabilities to $[10^{-7}, 1 - 10^{-7}]$, following the frozen human-evaluation metric routine; inner schedule selection uses $[10^{-6}, 1 - 10^{-6}]$.

Table 39: Content-held-out MuQ feedback adaptation on 2,397 pair-mean targets. ZERO/10D initialize from human-only/10D-supervised training; D counts teacher score inputs. Treatments average 25 model probabilities. Frozen retains original predictions; old data only replays original labels; temperature only fits a scalar; feedback + replay adds training-fold judgments to replay. NLL is negative log likelihood; soft Brier is squared error against the mean target (both lower-is-better). Accuracy is descriptive on 2,232 resolved directional majorities. Every fit excludes test-source components; listeners can contribute across folds.

Initialization	Treatment	NLL ↓	Soft Brier ↓	Accuracy (%)
ZERO start	Frozen	0.413764	0.055665	87.99
ZERO start	Old data only	0.434135	0.063271	86.29
ZERO start	Temperature only	0.407267	0.054069	87.99
ZERO start	Feedback + replay	0.383817	0.045503	89.52
10D start	Frozen	0.410388	0.054713	87.72
10D start	Old data only	0.427703	0.061278	86.38
10D start	Temperature only	0.405976	0.053691	87.72
10D start	Feedback + replay	0.381315	0.044751	89.78

All six adjusted tests favor new feedback (Table 40). Both starts improve over their frozen versions in each of the five content folds (Table 41) and in each seed's five-model ensemble. The two full-ensemble NLL reductions are approximately 7.24% and 7.08%. Descriptive majority accuracy rises from 87.99% to 89.52% for ZERO and from 87.72% to 89.78% for 10D. Accuracy uses 2,232 pairs with a resolved directional majority, gives tied predictions half credit, and excludes unresolved targets; no accuracy significance test is claimed. Soft-target Brier scores also decrease (Table 39). Old-data continuation increases NLL, whereas feedback improves over both the frozen and calibrated controls, so the reported advantage is not solely a consequence of deterioration in the continued training control. No test here establishes a ranking between the two adapted starts.

Verification and computational scope. Training used cached features for 2,368 unique new audio items and the existing judgments. Fits ran on CPUs, with two threads per outer-fold worker. CPU

Table 40: Content-held-out adaptation contrasts on 2,397 human pair-mean targets. Differences subtract each control's negative log likelihood (NLL) from feedback-plus-replay NLL; negative favors feedback. ZERO/10D denote human-only/10D-supervised MuQ starts; D counts teacher score inputs. Controls retain frozen predictions, train on original labels only, or fit a scalar temperature. Nominal 95% intervals use 20,000 source-component draws within fold, conditional on fitted predictions. p_{Holm} adjusts all six tests together.

Initialization	Control	ΔNLL	Conditional 95% CI	p_{Holm}
ZERO start	Old data only	-0.050318	$[-0.057704, -0.043293]$	0.0003
ZERO start	Frozen	-0.029947	$[-0.034943, -0.024929]$	0.0003
ZERO start	Temperature only	-0.023450	$[-0.028802, -0.018004]$	0.0003
10D start	Old data only	-0.046388	$[-0.053408, -0.039763]$	0.0003
10D start	Frozen	-0.029073	$[-0.033356, -0.024786]$	0.0003
10D start	Temperature only	-0.024661	$[-0.029395, -0.019979]$	0.0003

Table 41: Training-side schedule selection and held-out content-fold results for MuQ feedback adaptation. ZERO/10D are human-only/ 10D-supervised starts; D counts teacher score inputs. All schedules use learning rate 10^{-4} and feedback-loss weight $\alpha = 0.5$; remaining weight replays original labels. Negative ΔNLL (negative log likelihood) favors feedback over frozen predictions. T belongs to the separate temperature-only control and transforms its frozen ensemble probability as $\sigma(\mathrm{logit}(p)/T)$.

		ZERO start			10D start		
Fold	Test pairs	Epochs	T	ΔNLL	Epochs	T	ΔNLL
0	480	9	0.777063	-0.028647	9	0.807986	-0.026770
1	480	10	0.780418	-0.009139	6	0.815785	-0.014462
2	479	9	0.792100	-0.040173	9	0.828815	-0.036423
3	479	10	0.792102	-0.034241	10	0.824524	-0.031968
4	479	5	0.802404	-0.037580	5	0.837015	-0.035776

predictions from the 50 starting models reproduce archived frozen probabilities within 4.77×10^{-7}. The final audit checks all 500 adaptation artifacts, unchanged frozen parameters, target and fitting-scope bindings, and actual parameter updates. Twenty saved models were independently reloaded for prediction with zero observed difference; the 24 NLL/Brier/accuracy cells were independently recalculated with zero difference. These checks establish internal consistency, not external replication. Appendix D.6 describes the separate de-identified statistical replay, whose scope excludes model fitting.

D.15 Supplementary adaptation with held-out listeners and content

The primary adaptation analysis evaluates new content judged by the same annotator pool. This supplement asks whether feedback from a subset of that pool also improves prediction of judgments from excluded annotators on disjoint content. It uses the same pairwise training method, with a different separation of people and sources. The execution archive indexes it as T33.

Fixed reservation and targets. Among the 23 QC-accepted annotators, each was assigned two of 12 task packets. There are 34 six-person subsets that cover every packet exactly once. A fixed hash order with seed 20260916 selects one such subset using packet assignments alone, without ranking preferences, reliability, or model errors. These six people are excluded from all new feedback fitting, selection, and calibration; the remaining 17 form the development group. The content partition retains the first source-component fold of the preceding analysis, selected by index, and its inner fitting/validation split. Both the full annotation pool and preceding adaptation outcomes had already been inspected. Fixing this reservation before the new fits therefore does not make it prospective confirmation.

For development sources, targets are recomputed using only the 17 people's nonmissing judgments, with ties as 0.5. One validation pair has no usable vote from this group, leaving 1,533 inner-fit pairs and 383 validation pairs. Final fitting uses their union: 1,916 pairs, 456 source components, and

5,333 votes. Test targets use only judgments from the six reserved people. Seven of the 480 reserved pairs have no usable test-group vote and are excluded without replacement, leaving 473 pairs from 114 source components. Each has one valid judgment. The split is not redrawn to compensate for missing votes. The development group's judgments on test sources and the test group's judgments on development sources are both excluded from training. No source component, audio endpoint, or audio hash crosses fitting, validation, and test boundaries. Test audio hashes also exclude the original human replay data, and test components exclude pseudo-training sources.

Refitting and controls. We restart from the same 25 original pre-feedback checkpoints per initialization, retaining the original human-label replay scope for each checkpoint. None of the preceding adaptation's weights, selected schedules, temperatures, or all-person target means are reused. Frozen audio features are reused. The search space, trainable parameters, loss, and matched continuation control are those of Appendix D.14. Six inner trials per initialization select learning rate, feedback weight, and epoch using only the 17-person validation targets. Both starts select learning rate 10^{-4}, weight $\alpha = 0.5$, and 12 epochs, the fixed search limit. Fifty feedback fits and 50 old-data-only fits each take 360 optimizer steps. As before, continuation shares the feedback-selected schedule rather than optimizing a separate one.

One positive temperature per initialization is fitted on the 1,916 development targets, yielding $T = 0.769714$ for ZERO and $T = 0.797556$ for 10D. All final fits, schedules, and temperatures are sealed before computing test metrics. The optimizer target tensor contains missing values outside the development scope, including every test row. Each treatment averages 25 model probabilities.

Evaluation and complete contrasts. The primary endpoint is NLL on the 473 reserved single-judgment targets. We report all six feedback-minus-control contrasts, using one Holm family with level 0.05. The 20,000 paired source-component bootstrap draws use seed 330916. The 95% percentile intervals are nominal; they and the adjusted tests are conditional on the six reserved people, the fixed partition, and the fitted models. They do not include listener-population or full-retraining uncertainty. Metric clipping and tie handling follow the preceding adaptation analysis. Descriptive accuracy excludes the 48 tied judgments, leaving 425 directional targets.

Table 42: MuQ adaptation with held-out listeners and content: 17 people supply development feedback; six others supply 473 single-judgment test targets. ZERO/10D start from human-only/10D-supervised training; D counts teacher score inputs. Treatments average 25 model probabilities. Frozen retains original predictions; old data only replays original labels; temperature only fits a scalar; feedback + replay adds development judgments to replay. NLL is negative log likelihood; soft Brier is squared error, coding ties as 0.5 (both lower-is-better). Descriptive accuracy uses 425 non-tie judgments.

Initialization	Treatment	NLL ↓	Soft Brier ↓	Accuracy (%)
ZERO start	Frozen	0.467219	0.127592	82.12
ZERO start	Old data only	0.514285	0.139519	79.53
ZERO start	Temperature only	0.474898	0.128800	82.12
ZERO start	Feedback + replay	0.450289	0.120712	82.12
10D start	Frozen	0.465657	0.126921	82.12
10D start	Old data only	0.507591	0.136617	80.24
10D start	Temperature only	0.474516	0.128313	82.12
10D start	Feedback + replay	0.447054	0.119437	83.06

Table 42 shows lower mean NLL after feedback adaptation, with relative decreases of 3.62% for ZERO and 3.99% for 10D against their frozen predictions. Table 43 reports the complete adjusted comparisons separately from nominal intervals. All five source-seed ensembles have lower test NLL than their corresponding frozen ensembles. This seed-level pattern is a descriptive sensitivity, not an additional independent experiment.

This supplement evaluates a different target from the primary adaptation: single-listener judgments from excluded annotators, rather than the within-pair means of the original pool. Its losses and relative changes do not rank 17-person training against 23-person training. The reservation establishes person isolation within this adaptation cohort; identity overlap with older training studies has not been established. Six listeners do not constitute a new population sample, and 473 judgments are not 473

Table 43: All six feedback-minus-control comparisons on 473 judgments from six excluded listeners and disjoint sources. Differences use negative log likelihood (NLL); negative favors feedback plus original-label replay. ZERO/10D are human-only/10D-supervised MuQ starts; D counts teacher score inputs. Controls are frozen, original-label-only training, or scalar calibration. Nominal 95% intervals use 20,000 source-component bootstrap draws, conditional on these listeners and fits. p_{Holm} adjusts six tests together at level 0.05.

Initialization	Control	ΔNLL	Conditional 95% CI	p_{Holm}
ZERO start	Old data only	-0.063996	$[-0.094875, -0.036319]$	0.001500
ZERO start	Frozen	-0.016930	$[-0.033993, 0.000346]$	0.052847
ZERO start	Temperature only	-0.024608	$[-0.042471, -0.007225]$	0.021149
10D start	Old data only	-0.060537	$[-0.090082, -0.033505]$	0.001500
10D start	Frozen	-0.018603	$[-0.034986, -0.002176]$	0.051897
10D start	Temperature only	-0.027462	$[-0.045757, -0.009903]$	0.011599

independent listeners. Neither adaptation experiment measures downstream teaching or generated-music quality.

Packaged MuQ reward model and inference. The packaged audio-only MuQ reward model uses the human-only (ZERO) ensemble adapted in the listener-and-content holdout (Tables 42 and 43). It includes the shared frozen encoder and projection with all 25 adapted members. For audio a, b, it averages the $M = 25$ members' preference probabilities, $p(a \succ b) = M^{-1} \sum_m \sigma(s_m(a) - s_m(b))$, where s_m is member m's scalar score. The single-audio score $r(x) = M^{-1} \sum_m s_m(x)$ is an uncalibrated relative utility; generally $p(a \succ b) \neq \sigma(r(a) - r(b))$.

Verification and replay. An independent audit reconstructs all 2,389 development/test targets from the accepted per-person returns, verifies person and source isolation, and checks all 100 saved model states for parameter updates and frozen-parameter integrity. Eight checkpoint inference replays match exactly; 24 independently reconstructed NLL/Brier/accuracy values agree within 6×10^{-17}. The statistical package in Appendix D.6 reproduces fixed-prediction metrics and contrasts without repeating training, selection, or listening.

E Counterexample, Local Utility, and a Sufficient Regime

This appendix proves the NLL counterexample, derives the local utility identity, and gives a sufficient Brier-risk regime. The experiments use selected, weighted targets and AdamW training; their teaching utility is assessed empirically without assuming this regime.

E.1 Proof of the log-loss counterexample

Proposition 1 uses the same population, examples, unit weights, and soft-target interface for both evaluators.

Proof. Take three equally weighted examples with human labels $(1, 1, 0)$ and evaluator probabilities $p_{E_1} = (.6, .6, .1)$ and $p_{E_2} = (.55, .55, .60)$. Their direct NLLs are $0.3757 < 0.7040$. Let the constrained student predict the same sigmoid probability $\sigma(\theta)$ on every example, start at $\theta = 0$, and take one step of size η on mean soft-target cross-entropy; ZERO stays at $\theta = 0$. Since $\theta'_E = \eta(\bar{p}_E - 1/2)$ for mean teacher target $\bar{p}_E$, the means $13/30$ and $17/30$ give $\theta'_{E_1} = -\eta/15$ and $\theta'_{E_2} = \eta/15$. Human NLL is $R_H(\theta) = \log(1 + e^\theta) - \frac{2}{3}\theta$, whose derivative $\sigma(\theta) - 2/3$ is negative for $\theta < \log 2$. Thus E_1 increases human NLL for every $\eta > 0$, whereas E_2 decreases it whenever $0 < \eta < 15 \log 2$. □

E.2 Local utility and the supervision interface

Let $R_H(\theta)$ be the student's human risk after the shared part of a matched schedule, and let $R_{E,A}(\theta)$ be the mean weighted pseudo-risk before its scalar loss multiplier. For the local identity, absorb that

multiplier into η. For one additional pseudo step with $g_{E,A} = \nabla R_{E,A}(\theta)$, $\theta'_E = \theta - \eta g_{E,A}$, assume R_H is twice continuously differentiable in a neighborhood of the step segment. Taylor expansion gives

$$R_H(\theta) - R_H(\theta'_E) = \eta\langle \nabla R_H(\theta), g_{E,A}\rangle - \frac{\eta^2}{2} g_{E,A}^\top H_H(\widetilde{\theta}) g_{E,A}. \tag{10}$$

Here H_H is the Hessian of R_H, and $\widetilde{\theta}$ lies on the segment from θ to θ'_E. At fixed θ and pseudo-risk, bounded curvature along sufficiently short steps gives the $O(\eta^2)$ remainder in Equation 5 as $\eta \to 0$. The identity separates gradient alignment from curvature for a plain gradient step. The experiments use AdamW and are assessed through their finite training outcomes against schedule-matched ZERO.

The evaluator–interface boundary. Let $q_i^E \in \{0,1\}$ indicate whether A selects pseudo pair z_i, let $\hat{y}_i^E \in [0,1]$ and $w_i^E \geq 0$ be its induced target and weight, and let $Z_E = \sum_i q_i^E > 0$. At a fixed student state, the pseudo gradient is

$$g_{E,A}(\theta) = \frac{1}{Z_E} \sum_i q_i^E w_i^E \nabla_\theta \ell\big(S_\theta(z_i), \hat{y}_i^E\big)\,. \tag{11}$$

Here $S_\theta(z_i)$ is the student probability and $\ell(p,y)$ is the fractional log loss in Equation 3. Selection, targets, and weights are fixed when differentiating with respect to the student. The denominator counts selected examples; it does not normalize by their total confidence weight. Thus a richer evaluator can change the support, labels, or weights even when pair budget and optimizer are matched. Comparing teacher predictions alone does not isolate these distinct contributions to the student update.

Interface equivalence. With the student, frozen inputs, initialization, randomness, optimizer, and human schedule fixed, identical ordered pseudo examples, targets, weights, and pseudo-loss schedules imply the same trajectory: identical minibatches give identical gradients and optimizer states at each deterministic update. Equal pair budgets alone do not meet these conditions because selection, targets, weights, or pseudo-loss schedules can still differ.

E.3 A SUFFICIENT BRIER-RISK TRANSFER REGIME

Observation 1 (A sufficient Brier-risk transfer regime)**.** *Let $\mu \in [0,1]^n$ be human conditional preferences on n equally weighted examples and $f \in [0,1]^n$ the student's current predictions. Suppose A uses the same examples and unit weights for every evaluator, uses soft targets $p_E \in [0,1]^n$, and one prediction-space update is $f'_E = (1-a)f + ap_E$ for fixed $a \in (0,1)$. If evaluator residuals $\epsilon_E = p_E - \mu$ satisfy $\langle f - \mu, \epsilon_E\rangle = 0$ for each evaluator being compared, define $R_H^{(2)}(h) = \|h-\mu\|^2/(2n)$ and $U^{(2)}(E) = R_H^{(2)}(f) - R_H^{(2)}(f'_E)$. Then*

$$R_H^{(2)}(f'_E) = \frac{(1-a)^2}{2n}\|f-\mu\|^2 + \frac{a^2}{2n}\|\epsilon_E\|^2. \tag{12}$$

Hence lower direct Brier excess risk implies higher $U^{(2)}$. This is not the NLL utility in Equation 4. Here $R_H^{(2)}$ is half the Brier excess risk, omitting the human-label variance term. Empirical Brier scores use mean squared error without the factor $1/2$.

Proof. Writing $d = f - \mu$, expand

$$\|f'_E - \mu\|^2 = (1-a)^2\|d\|^2 + 2a(1-a)\langle d, \epsilon_E\rangle + a^2\|\epsilon_E\|^2.$$

The orthogonality assumption removes the middle term, giving Equation 12. Subtracting from the unchanged student risk yields

$$U^{(2)}(E) = \frac{a}{2n}\left[(2-a)\|f-\mu\|^2 - a\|\epsilon_E\|^2\right].$$

Thus utility decreases strictly with squared evaluator error under the shared assumptions. Positive utility additionally requires $a\|\epsilon_E\|^2 < (2-a)\|f-\mu\|^2$; the ranking implication alone does not ensure a gain over the unchanged student. □