Personalized Grounding of Music Concepts: Dialogue-Based Query Refinement and Residual Reranking

Yota Minoshima Shun Okuhara
Division of Information Engineering, Graduate School of Engineering
Mie University
Tsu, Japan
youta1123@icloud.com okuhara@eng.mie-u.ac.jp

***Abstract*— Interpreting subjective music requests requires connecting intended meanings with perceived qualities of sound. We propose personalized grounding (PG), which represents term–concept semantics separately from concept–audio models. Perceptual feedback updates the relevant acoustic model, competing hypotheses guide subsequent comparisons, and the learned state supplies a residual correction to query-based retrieval. This loop connects the evidence obtained, the model updated, and the next question asked. We compare Rewrite–CLaMP3, acoustic relevance feedback (Audio-RF), and Rewrite+PG in an exploratory closed-loop simulated-user study of 57 queries. Adaptation occurred in acoustic grounding while semantic weights retained their initial values. PG's mean final target-audio similarity on unexposed tracks was below both baselines, whereas its mean final request-fit rating was highest on the common 33-query subset. Request fit was rated in separate calls to the simulator's LLM, with dialogue history and method identity withheld. Paired between-method intervals included zero for both outcomes. In a post-hoc diagnostic holding each final PG history and query fixed, adding the final learned residual improved target-audio similarity on average. This conditional observation concerns the last ranking step; the full-dialogue comparison combines learning with different interaction policies. Dialogue cases linked gains to request-relevant acoustic descriptions and declines to narrow question focus and accumulated constraints. These findings distinguish the conditional contribution of reranking from retrieval outcomes along different dialogue trajectories. They identify question relevance, feedback scope, and the connection between perceived and desired attribute intensity as design priorities. Creativity support motivates the framework and remains a target for subsequent human evaluation.**

***Index Terms*—conversational music retrieval, personalization, concept grounding, query rewriting, user simulation**

I. Introduction

Conversational music recommendation offers opportunities for clarifying listening goals, exploring unfamiliar music, and reflecting on personal taste. Yun and Lim's study of LLM-powered recommenders describes implicit-need clarification, self-directed exploration, and deeper understanding of musical preferences [1]. These experiences motivate a creativity-support perspective: helping listeners formulate and revise meaningful exploration criteria through conversation and listening.

Recent systems connect conversation to music retrieval in several ways. TalkPlay learns to generate music tokens from dialogue and multimodal music representations [2]. TalkPlay-Tools coordinates text-to-audio search, example-based retrieval, metadata filtering, and listening-history-based personalization through LLM tool calls [3]. These approaches establish practical ways to translate conversational context into catalog operations. Our focus is the feedback-dependent interpretation used within the acoustic retrieval component.

Music requests contain preferences, exclusions, references, and contextual meanings [4], [5]. Consider a listener requesting heavier music. The expression may refer to stronger bass, a rougher timbre, or denser rhythmic activity. After choosing timbral roughness as an interpretation, the system must also determine which recordings express that quality for this listener. A statement about what heavier means and a judgment about how rough a recording sounds address different uncertainties.

Query rewriting incorporates feedback into a search description, while audio relevance feedback changes a query vector using overall request-fit comparisons. PG instead assigns evidence to explicit learning targets: judgments about a named audible property update that concept's acoustic direction g_k, and supported statements about a term's meaning update its concept weights π. This builds on personalized soft-attribute semantics [6]. Its organizing principle is the connection among the evidence obtained, the concept model revised, and the next comparison offered to the listener.

The acoustic hypotheses guide track comparisons and concept-specific questions. Their public replies enter the shared rewriter, while the updated grounding model provides an initial-state-relative ranking correction. A perceptual judgment can therefore influence later dialogue and unpresented-track ranking. The design aims to make feedback actionable at the concept level. Its retrieval value depends on whether the selected concepts express desirable qualities for the current request, whether the feedback supports the acoustic update, and whether the next comparison addresses the remaining uncertainty.

We compare Rewrite–CLaMP3, Audio-RF, and Rewrite+PG in a paired, closed-loop simulated-user study of 57 queries. The methodological contribution is an evidence-typed loop that links concept-specific acoustic learning to the next comparison and to residual reranking. The empirical contribution is an

analysis of g-centered adaptation under unchanged semantic weights, combining between-system comparisons, a final-correction diagnostic conditioned on saved histories, and contrasting dialogue cases. These analyses identify priorities for request-relevant questioning and feedback interpretation. Creativity support motivates this design and remains a target for subsequent human evaluation.

II. Related Work

Conversational music retrieval. TalkPlay and TalkPlay-Tools represent generative and tool-based approaches to music recommendation [2], [3]. The latter combines multiple sources of evidence, including text–audio similarity and user listening history. PG addresses a complementary level: revising subjective term and sound interpretations from within-session feedback. We use a controlled rewrite-and-retrieve pipeline with CLaMP 3 [7] as our comparator. The experiment isolates this retrieval layer rather than benchmarking the complete TalkPlay agents.

Personalized semantics and elicitation. Göpfert et al. model personalized meanings of soft attributes with concept activation vectors, including variation in attribute degree and meaning [6]. Biyik et al. combine item and soft-attribute queries for preference elicitation under semantic uncertainty [8]. PG brings these ideas into a music dialogue with an LLM rewriter. Its emphasis is on routing semantic and perceptual evidence to distinct states, revising the relevant acoustic models from presented recordings, and using disagreement among acoustic hypotheses to propose the next comparison. The study examines this combination with residual reranking.

User experience and creativity support. Yun and Lim conducted a three-week diary study with 12 participants using customized GPT-based music recommenders [1]. Participants described explaining their tastes through musical attributes and using familiar preferences to explore new genres. The study also identified dialogue burden and persuasive explanations that could constrain reflection. These findings motivate computational support for articulating and revising exploration criteria. Our simulation examines the feedback and retrieval mechanisms underlying that design goal; human experience is a subsequent evaluation target.

Data and evaluation. MusiCRS compares audio, query, and combined inputs for music recommendation [9]; Reddit2Deezer supplies request–item associations and paraphrased conversations [10]. MusicSem documents a broad range of musical meanings [5], while generative-user research treats simulator fidelity as an empirical question [11]. Together they motivate evaluating acoustic retrieval, request fit, and dialogue interpretation as distinct outcomes.

III. Personalized Music Concept Grounding

Figure 1 summarizes personalized grounding. Panel (a) separates term-to-concept semantics π from concept-to-audio grounding g through a hypothetical request for heavier music. A clarification of the intended meaning informs π, whereas a judgment that X sounds rougher than Y informs g together with the presented tracks' audio embeddings z_X and z_Y.

TABLE I
Illustrative feedback types and their update targets.

Evidence	Example	Target
Meaning	By heavier, I mean rougher timbre.	π
Perception	X sounds rougher than Y.	g
Global fit	X fits my request better.	h (RF)
Preference	I like X more.	Separate channel

Panel (b) shows how the learned state acts through two pathways. In pathway A (dashed), g and its hypotheses guide the next comparison and question; the listener's public reply enters the history used by the shared rewriter. In pathway B (solid), the learned π/g state supplies a residual ranking correction relative to its initialization. The encoders remain fixed, while component directions and posterior weights adapt within the fixed hypothesis pool.

A. Semantic and acoustic states

For an expression in request q, π_k weights candidate concept k, and $g_k(x) = c_k^\top z_x$ predicts its realization in track x, with normalized audio embedding z_x. For example, a correction from bass to roughness changes the semantic interpretation, whereas a comparison of perceived roughness informs the acoustic model. Concepts and candidate acoustic descriptions are initialized from the public request. Term semantics use a symmetric Dirichlet prior ($\alpha_k = 1$), with a session-working state and a conservative confirmed state.

Each concept has J_k candidate descriptions. The frozen text encoder maps description j to prior direction $b_{k,j}$. We retain competing acoustic interpretations through a Gaussian mixture prior:

$$p(c_k) = \frac{1}{J_k} \sum_{j=1}^{J_k} \mathcal{N}(c_k; b_{k,j}, \tau^2 \mathbf{I}). \tag{1}$$

Feedback adapts component directions and weights within the fixed hypothesis set.

B. Evidence-specific updates

An evidence router decomposes each response into atomic observations (Table I), identifying the evidence type, concept, presented track or pair, and response. A perceptual atom about concept k refits only that concept's acoustic components. Supported term definitions, term–concept relations, and eligible concept-scoped request-fit observations target π. Global fit supplies Audio-RF's separate update channel. The LLM extracts this structure; likelihoods and priors determine the numerical adaptation.

Let $s(t) = (1 + \exp(-t))^{-1}$. For an absolute perceptual observation about track x, we use the concept's mean μ_k^{pre} and scale σ_k^{pre} on C immediately before its parent utterance. The probability of a positive attribute judgment is

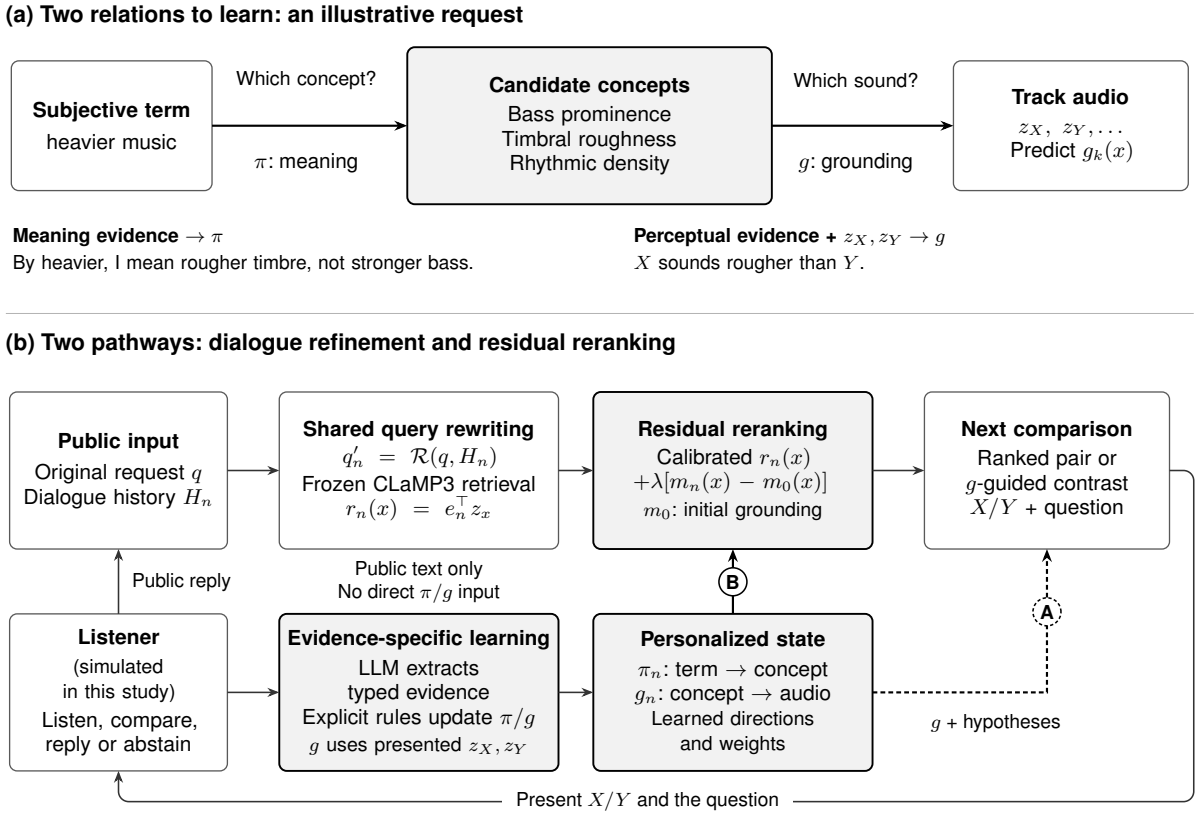


Fig. 1. Overview of personalized music concept grounding.

$$P(y = + \mid c, x) = s\left(\frac{(c^\top z_x - \mu_k^{\text{pre}})/\sigma_k^{\text{pre}} - \theta}{T_e}\right). \quad (2)$$

The negative probability is its complement. For a pair, $f = c^\top(z_Y - z_X)$ and the outcomes X, Y, and tie have probabilities P_X, P_Y, and P_{tie}:

$$\begin{aligned} P_Y(c) &= s((f - \varepsilon)/T_e), \\ P_X(c) &= s((-f - \varepsilon)/T_e), \\ P_{\text{tie}}(c) &= 1 - P_Y(c) - P_X(c). \end{aligned} \quad (3)$$

We set $\theta = 0$ and $\varepsilon = 0.10$. The evaluated PG condition uses a fixed likelihood temperature $T_e = 1.0$ for all accepted perceptual observations. Atoms share their parent utterance's pre-update calibration and retain it during subsequent refitting. For accepted observations $\mathcal{O}_{k,n}$, component j minimizes

$$\begin{aligned} \hat{c}_{n,k,j} &= \arg\min_c \ \mathcal{L}_{n,k,j}(c), \\ \mathcal{L}_{n,k,j}(c) &= \frac{\|c - b_{k,j}\|_2^2}{2\tau^2} - \sum_{e \in \mathcal{O}_{k,n}} \log P(y_e \mid c, e). \end{aligned} \quad (4)$$

L-BFGS-B fits $\hat{c}_{n,k,j}$. Diagonal Laplace approximation gives covariance $\Sigma_{n,k,j}$ and evidence $Z_{n,k,j}$, with normalized weights $\omega_{n,k,j} = Z_{n,k,j} / \sum_l Z_{n,k,l}$. The observations therefore revise both acoustic directions and the support for competing descriptions. Perceptual prediction accuracy is a separate evaluation target.

The semantic channel uses the prior mean $p_k = E[\pi_k]$, concentration $A = \sum_k \alpha_k$, and concept-specific likelihood ℓ_k:

$$p'_k = \frac{p_k \ell_k^{\rho_L}}{\sum_i p_i \ell_i^{\rho_L}}, \qquad \alpha'_k = (A + \nu_L) p'_k. \quad (5)$$

The evidence level sets (ρ_L, ν_L) to $(0, 0)$, $(0.25, 0.25)$, $(1, 1)$, or $(1, 4)$, giving explicit definitions stronger weight than implicit evidence. Confirmed semantics use a separate promotion rule. The present study exercises the acoustic-learning channel: both semantic states retain their initial values in all 57 queries.

C. Residual reranking

A shared rewriting procedure forms $q'_n = \mathcal{R}(q, H_n)$ from the original request and the method's public dialogue history. With frozen encoders, the base retrieval score is

$$\begin{aligned} e_n &= \frac{E_{\text{text}}(q'_n)}{\|E_{\text{text}}(q'_n)\|_2}, \quad z_x = \frac{E_{\text{audio}}(x)}{\|E_{\text{audio}}(x)\|_2}, \\ r_n(x) &= e_n^\top z_x. \end{aligned} \quad (6)$$

Component MAP directions $\hat{c}$ and posterior weights ω yield a model-averaged concept score. We standardize each concept

on a fixed calibration pool C and combine it with the working semantic weights:

$$\bar{g}_{n,k}(x) = \sum_j \omega_{n,k,j} \hat{c}_{n,k,j}^\top z_x, \tag{7}$$

$$\begin{aligned} \tilde{g}_{n,k}(x) &= \frac{\bar{g}_{n,k}(x) - \mu^g_{C,n,k}}{\sigma^g_{C,n,k}}, \\ m_n(x) &= \sum_k E[\pi^{\text{working}}_{n,k}] \tilde{g}_{n,k}(x). \end{aligned} \tag{8}$$

Let m_0 retain the initial grounding model and calibration. The proposed score adds its change to the standardized rewrite score:

$$S_{\text{PG},n}(x) = \frac{r_n(x) - \mu^r_{C,n}}{\sigma^r_{C,n}} + \lambda[m_n(x) - m_0(x)]. \tag{9}$$

The calibration identities are fixed, and their statistics follow the current model state. The residual begins at zero, preserving the baseline's initial ranking. Its magnitude reflects both the learned change and λ. The broader request remains represented by the rewritten query, while the residual expresses adaptation of the concept–audio interpretation.

D. Connecting perception to the request

Desirability enters through request-derived concepts and the nonnegative mixture in Eq. (8). The scorer assumes that stronger realization of a candidate meaning supports the requested expression. Here π weights interpretations, while rewriting carries broader requirements and exclusions. The evaluated scorer has no separate learned parameters for desired intensity or preference polarity. The initial-state subtraction preserves this coupling assumption.

For example, X sounds rougher but is too rough for this request contains a perceptual observation for g and a desired-level constraint for rewriting. With π fixed here, $m_n - m_0 = \sum_k p_{0,k}(\tilde{g}_{n,k} - \tilde{g}_{0,k})$ is a positively weighted acoustic-model change. A separate representation of whether an attribute should increase, decrease, or reach an intermediate level would extend this request–perception connection.

E. Concept-directed comparisons

PG alternates ranked recommendations with concept-directed comparisons. For comparison turns, eligible tracks come from the current raw rewrite's top 1,000 in I. For each concept, the bounded candidate search combines up to 4,096 uniformly sampled pairs with pairs formed from the eight highest- and eight lowest-scoring tracks under each component. Let $p_j(y \mid X, Y)$ be component j's predictive distribution over the three outcomes in Eq. (3). Pair selection maximizes the following disagreement proxy within this candidate inventory:

$$A_k(X,Y) = H\left(\sum_j \omega_{n,k,j} p_j\right) - \sum_j \omega_{n,k,j} H(p_j). \tag{10}$$

H is categorical entropy. To include diagonal posterior uncertainty, let $d = z_Y - z_X$ and $a_j \simeq \sqrt{1 + 0.393\, d^\top \Sigma_{n,k,j} d}$, where $0.393 \approx \pi/8$ arises from a probit approximation to the logistic function [15]; component prediction divides the pairwise logits by a_j. This proxy seeks disagreement between hypotheses, targeting perceptual discrimination rather than immediate request fit. The LLM chooses a concept from these options using the request and public history. The selected pair and concept explanation form the question. Pair discrimination and request-relevant concept selection are thus separate decisions.

These pathways interact over turns: ranking and grounding affect subsequent comparisons, replies, and queries. The closed loop evaluates their combined operation; the saved-history diagnostic examines the final correction for a fixed dialogue-conditioned query.

IV. Experimental Design

A. Queries and track pools

We analyzed 57 source-query clusters, each with three ten-turn conditions (171 sessions and 1,710 turn slots). A cluster corresponds to a Reddit thread and leaf-comment branch, and is the unit of analysis.

New dialogues start from Reddit2Deezer paraphrased requests [10]. A deterministic rule selects the lexicographically first structurally positive linked track as private target G_q. Such a link supplies a reference recording associated with a recommendation. The study operationalizes an open-ended request using that fixed recording. Interaction pool I contains 50,000 tracks, unexposed evaluation pool T contains 5,000, and calibration pool C contains 1,000. The pools and private targets are disjoint. The system accesses public replies and presented-track embeddings; targets remain on the simulator and evaluation side.

B. Compared systems and simulated feedback

Rewrite–CLaMP3 ranks by Eq. (6) and presents its two highest-ranked available, unpresented tracks. Audio-RF uses the same policy with an additional audio residual h_n. Given an eligible public judgment that x^+ fits the whole request better than x^-, it updates and scores as follows:

$$\begin{aligned} h_{n+1} &= 0.9h_n + 0.1w_n(z_{x^+} - z_{x^-}), \quad h_0 = 0, \\ S_{\text{RF},n}(x) &= \left(\frac{e_n + h_n}{\|e_n + h_n\|_2}\right)^\top z_x. \end{aligned} \tag{11}$$

High, medium, and low confidence map to $w_n = 1, 0.5, 0$. Eligible high/medium comparisons trigger the update; other responses retain h_n. PG uses $\lambda = 0.5$, three concepts with three hypotheses each, and $\tau = 0.25$. Odd turns present ranked pairs and even turns present concept-directed contrasts. Each method rewrites its own public history using the same procedure. Weights of the foundation models remain frozen, and each query–condition pair has one trajectory.

Qwen3-Omni-30B-A3B-Instruct's Thinker [12] generates feedback in two stages: private acoustic judgments relative

TABLE II
MODEL IDENTITIES AND EXECUTION SETTINGS.

Component	Recorded identity and settings
CLaMP3	`sander-wood/clamp3`, SAAS, 768D. Code revision: `9016d2b0c8d1`; weight SHA-256 prefix: `5033f868e397`.
Backbones	`m-a-p/MERT-v1-95M`, revision `12af15fef9d0`, 24 kHz; `FacebookAI/xlm-roberta-base`, revision `e73636d4f797`.
Text LLM	`Qwen/Qwen3.8-27B`, 27.3B; `qwen38_2:latest`.
GGUF	Unsloth UD-Q8_K_XL (Q8_0-dominant mixed precision). API file-level label: Q4_K_M.
Retrieval source	`unsloth/Qwen3.8-27B-GGUF`; retrieval revision: `4ca720788d1e`.
API digest	`a3fa43731e63cb2a0183ec4f47385795` `0e6748aa24b2195de20fd871ee8b761c`
Text settings	Ollama 0.32.15; context 32,768. Rewrite/router temperature 0; hypothesis proposal 0.6.
Audio LLM	`Qwen/Qwen3-Omni-30B-A3B-Instruct` (Thinker), revision `26291f793822`; BF16, SDPA. Deterministic decoding; 16 kHz mono input.

to G_q, followed by a public reply conditioned on audio, the original request, and the current question. The simulator receives only the current comparison context; the recommender accumulates the full public history. The simulator may describe desired acoustic features, differences, and uncertainty in one or two English sentences. Audio consists of 30-second previews. Separate representations for a private listening reference and a public request make this a target-conditioned simulation.

Table II identifies the evaluated services. Retrieval and embedding-based evaluation use frozen CLaMP3-SAAS with 768-dimensional normalized vectors and MERT-v1-95M features [7], [13]. The text service uses Qwen/Qwen3.8-27B [14], distributed as Unsloth's UD-Q8_K_XL GGUF and identified by its archived runtime alias and digest. Rewriting and routing use temperature 0, and hypothesis proposal uses 0.6. Methods share the initial query and ranking. The rewriter uses the request and public dialogue; acoustic states guide comparisons and residual ranking.

Recorded language-model and projector blob identifiers match the published SHA-256 identifiers of Unsloth's UD-Q8_K_XL GGUF and `mmproj-BF16.gguf` at Table II's retrieval revision. Tensor descriptors show Q8_0-dominant mixed precision and no four-bit tensors; Q4_K_M is Ollama's file-level label.

C. Outcomes and analysis

RQ1 compares retrieval before and after dialogue and between methods. The primary outcome is target-audio cosine@5 on T at turn 10. For candidate set $\mathcal{A}$, it is

$$V_{q,m,n}(\mathcal{A}) = \frac{1}{5} \sum_{x \in \mathrm{Top5}(S_{q,m,n}, \mathcal{A})} z_{G_q}^{\top} z_x. \tag{12}$$

TABLE III
WITHIN-METHOD CHANGES IN T COSINE@5 ($N = 57$).

Method	Final	$\Delta_{0\to10}$	95% interval	$+/-/=$
Rewrite	0.5168	$+0.0176$	$[-0.0036, +0.0386]$	36/20/1
Audio-RF	0.5261	$+0.0268$	$[+0.0042, +0.0492]$	38/18/1
Rewrite+PG	0.5097	$+0.0104$	$[-0.0168, +0.0371]$	31/25/1

Higher values indicate greater similarity of retrieved tracks to the reference in the fixed audio space. We also inspect Top1 cosine and independently rerank I and $I \cup T$. Checkpoints at turns 0, 1, 2, 4, 6, 8, and 10 describe the trajectory. RQ2 holds the final PG history and query fixed and removes its last residual term, locating the conditional ranking effect. RQ3 links questions, replies, query changes, and retained learning observations in the saved dialogues.

A separate-call Top1 evaluation uses the same Qwen3-Omni model to assign an integer rating from 1 to 10 or an unjudgeable response. Acoustic similarity supplies the reference and retrieved audio; request fit additionally supplies the original query. The reference is presented as A and the candidate as B. The evaluator receives the saved initial or final Top1, with dialogue history, method identity, rewritten query, learned state, embeddings, and cosine scores withheld. Identical inputs are deduplicated within the run. We compare methods on common numeric initial–final pairs for each metric (45 queries for acoustic similarity and 33 for request fit on T); missing and unjudgeable outputs remain excluded from these paired sets. These ratings are computed separately from CLaMP3 similarity, while sharing the simulator's model and reference conditioning. They provide complementary model-based evidence rather than an independent human criterion.

Queries receive equal weight. We report paired differences and descriptive 95% percentile bootstrap intervals, resampling queries 5,000 times. The saved-history correction diagnostic uses 10,000 resamples. The cohort had been used during development, and the intervals summarize exploratory uncertainty across several readouts. Missing or unjudgeable LLM scores retain their status and each table states its effective sample size. Fourteen queries were reviewed for contrasting changes and outcomes; selected examples illustrate the mechanisms in Section VI.

V. RESULTS

A. Retrieval outcomes

Mean target-audio cosine@5 increased in all three methods, with Audio-RF showing the largest gain, followed by Rewrite and PG (Table III). Table IV reports their final means. The mean paired difference for PG relative to Rewrite was -0.0071 (95% interval $[-0.0317, +0.0185]$); intervals for all paired method comparisons included zero. In Table III, $\Delta_{0\to10}$ denotes change from pre-dialogue retrieval, and $+/-/=$ counts improvements, declines, and ties.

On the common 33-query subset, PG had the highest mean final LLM request-fit rating (Table IV), whereas mean changes

TABLE IV
FINAL SCORES BY METHOD.

Method	T cosine@5 ($N = 57$)	LLM request fit 1–10 ($N = 33$)
Rewrite–CLaMP3	0.5168	7.303
Audio-RF	0.5261	7.424
Rewrite+PG	0.5097	7.545

TABLE V
SEPARATE-CALL LLM RATINGS OF T TOP1.

Method	Final	$\Delta_{0\to10}$	$+/-/=$
Acoustic similarity ($N = 45$)			
Rewrite	2.111	$+0.089$	10/8/27
Audio-RF	2.178	$+0.156$	12/8/25
Rewrite+PG	2.044	$+0.022$	12/12/21
Request fit ($N = 33$)			
Rewrite	7.303	$+0.364$	5/2/26
Audio-RF	7.424	$+0.485$	6/3/24
Rewrite+PG	7.545	$+0.606$	10/7/16

in LLM acoustic-similarity ratings were small (Table V). Request-fit improvement for PG relative to Rewrite was $+0.242$ (95% interval $[-0.333, +0.909]$). All paired request-fit intervals included zero. These ratings complement cosine@5, with each outcome describing a different aspect of retrieval on its own effective sample.

The ordering of mean cosine@5 gains persisted in the post-hoc subset of 49 queries after excluding requests with clear dependence on lyrics or album artwork. This sensitivity check directs attention to dialogue and ranking mechanisms alongside request modality.

B. Final correction and query trajectory

The final PG correction increased target-audio similarity for the saved queries on average. To examine this conditional comparison, let $V_{\mathrm{PG}}^{\mathrm{off}}$ denote the evaluation obtained by retaining each final PG history and query and setting $\lambda = 0$ only for the final ranking. The change from the common initial result V_0 can be written as

$$V_{\mathrm{PG}} - V_0 = (V_{\mathrm{PG}}^{\mathrm{off}} - V_0) + (V_{\mathrm{PG}} - V_{\mathrm{PG}}^{\mathrm{off}}). \tag{13}$$

The mean correction gain was $+0.0104$ (95% interval $[0.0017, 0.0193]$), with both improvements and declines across queries. The saved histories already embody earlier questions, selections, and corrections; Eq. (13) therefore describes a conditional comparison at the final ranking stage.

Retrieval using the final PG queries alone had a lower mean than retrieval using queries from the separately generated Rewrite trajectories. The final correction partly offset that gap. This contrast separates improvement at the final ranking stage from performance of the complete dialogue system.

C. Dialogue and learned-state changes

Observed adaptation was centered on g: acoustic models changed in 53 queries, while both working and confirmed π retained their initial values. The accepted evidence consisted mainly of absolute perceptual judgments, supplemented by pairwise comparisons. Table VI summarizes final query length, mean revisions over ten opportunities, and turns with state changes.

TABLE VI
DIALOGUE AND UPDATE SUMMARIES.

Method	Final words	Revisions	g	h
Rewrite	27.9	7.28	0	0
Audio-RF	28.1	7.60	0	65
Rewrite+PG	33.3	8.26	219	0

TABLE VII
SELECTED PG CASES ON T COSINE@5.

	Request	Query-only	PG	$\Delta_{0\to10}$	Corr.
Q127	Emotional intensity	0.619	0.724	$+0.231$	$+0.105$
Q125	Recent favorites	0.393	0.379	-0.172	-0.014
Q134	Prominent saxophone	0.403	0.403	$+0.000$	$+0.000$
Q150	Trap	0.598	0.622	-0.145	$+0.024$
Q161	Ice imagery	0.410	0.416	$+0.131$	$+0.005$
Q138	Fire imagery	0.519	0.477	$+0.094$	-0.042

PG produced longer final queries and more revisions than the baselines (Table VI). No apparent monotonic association was observed between its word-count increase and cosine@5 gain (Spearman $\rho = 0.011$). Description length characterizes the amount of text; the cases in Section VI examine whether the added descriptions address the request.

Comparison questions were concentrated on few concepts. In 50 of 57 queries, all contrast turns addressed the same concept. The top-1,000 candidate restriction retained a request-based search region, while the acquisition proxy prioritized perceptual discrimination within it. These observations motivate separate assessment of pair selection and dialogue-level changes in question focus.

Public replies remained available to rewriting when validation deferred a numerical update. Query refinement and learned-state changes therefore followed distinct evidence paths. The next section traces these paths through contrasting dialogue cases.

VI. HOW DIALOGUE CHANGES THE SEARCH

Contrasting cases trace the request, questions, replies, and rewritten queries (Table VII). Query-only keeps the final PG history and removes only the final correction; $\Delta_{0\to10}$ is the total PG change from pre-dialogue retrieval, and Corr. is the final correction's contribution. Values are rounded independently. Audio descriptions are attributed to the simulated listener. Quotations retain the recorded wording, and query summaries condense the original text.

A. Actionable acoustic specification

For Q127, a request for emotionally intense music, PG repeatedly asked about forceful delivery. At turn 2, the listener replied, *Y has a more forceful vocal delivery that matches what*

I'm looking for. At turn 7, it described aggressive vocals and driving rhythm as expressing the intended energy and emotional weight. The final query combined forceful delivery, heavy instrumentation, driving rhythms, and urgency. T cosine@5 rose from 0.493 to 0.724. The final query alone scored 0.619; the correction added 0.105. Rewrite ended at 0.660. Separate-call acoustic similarity rose from 1 to 8, while final request fit was unjudgeable. This case connects perceptual descriptions to gains in both query-based ranking and residual reranking.

Acoustic specification also occurred with the baseline. For Q103's atmospheric winter request, Rewrite incorporated quiet introspection, soft vocals, and piano, reaching 0.729 compared with PG's 0.706. Thus the added value of PG depends on which interpretations its questions and acoustic state contribute beyond feedback already available to the shared rewriter.

B. Question focus and unresolved feedback

Q125 asked for recommendations based on recent favorite tracks, with the actual list absent from the public request. PG asked about rhythmic drive on all five contrast turns. The turn-3 reply was unavailable. At turn 9, the listener reported that the candidates' energy and vocal styles differed from its listening cue. The final query specified a strong, steady, insistent pulse, while referring generically to a vocal style consistent with recent history. T cosine@5 fell from 0.550 to 0.379; the final query without correction scored 0.393. The saved query already accounts for most of the decline. The dialogue suggests a missed opportunity to ask about the repeatedly mentioned voice and energy differences.

Q134 requested non-jazz music with a standout saxophone part. Initialization selected vibe as the target expression and instantiated three concepts: melancholic melody, reverberant space, and restrained energy. Some g hypotheses mentioned saxophone, while the concept set organized comparisons around melody, reverberation, and energy rather than saxophone presence or prominence. PG asked about melancholic melody four times and restrained energy once. At turn 4, the listener stated: *The melancholic melody axis seems to be a different focus from the core request about the saxophone element.* This correction was retained in structured extraction and in the comparison selector's dialogue history. Subsequent contrast actions remained restricted to the existing three concepts. The final query retained saxophone and added melancholic qualities; its total cosine@5 change was approximately zero, and separate-call request fit remained 1.

This case locates a limitation after the correction has reached the system: the available comparison actions provide limited scope for acting on it. It motivates retaining explicit request constraints separately from word-meaning hypotheses and translating corrections about question relevance into dialogue-policy decisions. Improving extraction and history delivery alone leaves this action-space restriction in place.

C. Preserving useful intermediate descriptions

Figure 2 plots target-audio cosine@5 on the unexposed pool T for Q150, a concise trap request. Markers denote saved checkpoints, and lines connect successive observations for each method. The first reply identified a strong beat and rap style, and all three methods rose from 0.766 to 0.839. Later trajectories diverged as search descriptions changed. In PG, the query accumulated rhythmic detail and exclusions. At turn 6, the listener requested a more grounded, head-nodding trap beat. The final query emphasized an insistent pulse and excluded overly electronic, chaotic, or purely melodic styles, with cosine@5 reaching 0.622. The final correction added 0.024 to the query-only score of 0.598. The useful early result and later decline motivate distinguishing durable requirements from dissatisfaction with individual presented tracks.

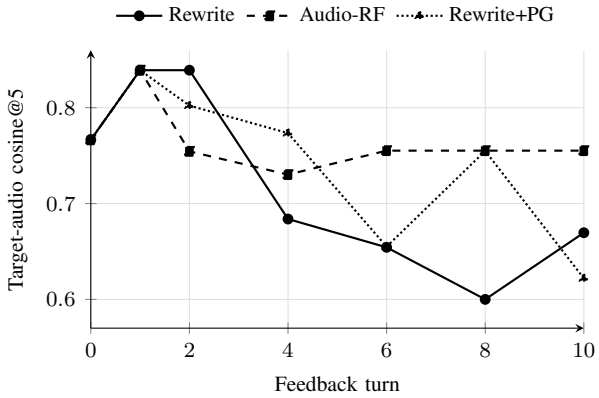


Fig. 2. Retrieval trajectory for Q150 (trap).

D. Acoustic resemblance and request fit

Q161 asked for the feeling of ice. PG specified restrained energy, gentle delivery, and a cool, quiet quality. Cosine@5 rose from 0.285 to 0.416 and separate-call request fit rose from 1 to 7, while acoustic similarity stayed at 1. In Q138's fire request, all methods raised request fit from 1 to 7; Rewrite emphasized driving energy and aggressive vocals, while PG emphasized rough texture. The corresponding final cosine@5 values were 0.554 and 0.477. These comparisons illustrate that fitting an image or mood can coexist with different degrees of resemblance to one fixed recording.

VII. Discussion and Design Implications

A. Matching the next question to the unresolved requirement

Question relevance provides a concrete design target. Repeated forceful-delivery questions supported Q127, whereas Q125's rhythmic questions left vocal-style feedback under-specified. Q134 distinguishes retaining a correction from offering a suitable response action. A policy could maintain explicit constraints alongside acoustic hypotheses, identify the remaining mismatch, and choose to continue a contrast, create a request-specific comparison objective, or ask for missing information. Saxophone presence and prominence are candidate objectives for Q134. Correction-following, subsequent retrieval, and later perceptual prediction can evaluate these design hypotheses.

B. Preserving the scope of feedback

Q150 motivates representing feedback as a requirement with scope: a judgment about this track, a preference for this session, or a persistent interpretation. A rewriter could retain accepted requirements, revise contradicted ones, and preserve uncertainty explicitly before producing a compact search description. Comparisons against the existing rewriter should measure requirement retention, unsupported constraint additions, and retrieval change.

Section III-D also motivates separating g_k from a request-specific instruction to increase, decrease, or target an intermediate attribute level. Strong perceived roughness and a wish for less roughness could then update distinct states. With matched evidence and candidate pools, request-fit and conditional reranking comparisons can evaluate this extension. The current positive-mixture scorer supplies the reference condition.

C. Supporting musical exploration

The creativity-support goal is to help listeners develop and reconsider their own exploration criteria. A comparison can make an abstract expression discussable through audible examples, as in forceful delivery or an icy atmosphere. Yun and Lim's user-experience findings [1] motivate evaluating this process through user-endorsed descriptions, perceived control, dialogue effort, and reuse of criteria on new music. The present traces identify candidate mechanisms for such support. A human study should examine whether listeners recognize the refined criteria as their own and can readily redirect the system when its interpretation diverges.

D. Scope of the evidence

The findings concern single-session adaptation to a fixed target with a memoryless simulated listener. They are conditioned on short previews, the reused available-case query sample, one trajectory per condition, and shared retrieval/evaluation embeddings. Cosine@5 measures proximity to one reference in the retrieval embedding space. Separate-call ratings provide a complementary view of request fit, while shared model and reference conditioning can produce correlated errors and the common numeric subset limits the requests represented. Separate-call LLM ratings use the simulator's model family, so human perceptual agreement and user benefit remain future evaluation targets. In this run, evidence supported changes to g, while π retained its prior, making elicitation of semantic evidence another concrete research target.

VIII. Conclusion

We proposed an evidence-typed music-grounding loop that connects concept-specific perceptual feedback, acoustic-model updates, and subsequent comparison questions, with a residual correction to query-based retrieval. In 57 simulated-user queries, adaptation occurred in g under unchanged π. The full-system PG–Rewrite difference was -0.0071, while adding the final PG residual to its saved query raised cosine@5 by 0.0104 on average. These results locate a conditional ranking contribution alongside variation in dialogue-conditioned queries. Contrasting cases identify request-relevant concept selection, feedback scope, and the separation of perceived intensity from desired intensity as next design priorities. Human evaluation of exploration and creativity support follows from this computational study.

References

[1] S. Yun and Y.-K. Lim, "User Experience with LLM-powered Conversational Recommendation Systems: A Case of Music Recommendation," Proc. CHI, Art.898, pp.1–15, 2025. doi:10.1145/3706598.3713347.

[2] S. Doh, K. Choi, and J. Nam, "TalkPlay: Multimodal Music Recommendation with Large Language Models," arXiv:2502.13713v5, 2025, revised 2026.

[3] S. Doh, K. Choi, and J. Nam, "TalkPlay-Tools: Conversational Music Recommendation with LLM Tool Calling," arXiv:2510.01698v4, 2025, revised 2026.

[4] M. Baranes, R. Hennequin, and E. V. Epure, "Beyond Musical Descriptors: Extracting Preference-Bearing Intent in Music Queries," Proc. NLP4MusA, pp.20–26, 2026. doi:10.18653/v1/2026.nlp4musa-1.4.

[5] R. Salganik et al., "MusicSem: A Semantically Rich Language–Audio Dataset of Natural Music Descriptions," arXiv:2602.17769, 2026.

[6] C. Göpfert et al., "Discovering Personalized Semantics for Soft Attributes in Recommender Systems Using Concept Activation Vectors," ACM Transactions on Recommender Systems, 2024. doi:10.1145/3658675. Author manuscript: arXiv:2202.02830.

[7] S. Wu et al., "CLaMP 3: Universal Music Information Retrieval Across Unaligned Modalities and Unseen Languages," Findings of ACL, pp.2605–2625, 2025. doi:10.18653/v1/2025.findings-acl.133.

[8] E. Biyik, F. Yao, Y. Chow, A. Haig, C.-w. Hsu, M. Ghavamzadeh, and C. Boutilier, "Preference Elicitation with Soft Attributes in Interactive Recommendation," arXiv:2311.02085, 2023.

[9] R. Surana, A. Namburi, G. Mundada, A. Lal, Z. Novack, J. McAuley, and J. Wu, "MusiCRS: Benchmarking Audio-Centric Conversational Recommendation," arXiv:2509.19469, 2025, revised 2026.

[10] H. Kim and J. McAuley, "Reddit2Deezer: A Scalable Dataset for Real-World Grounded Conversational Music Recommendation," arXiv:2605.09120, 2026.

[11] S.-e. Yoon, Z. He, J. Echterhoff, and J. McAuley, "Evaluating Large Language Models as Generative User Simulators for Conversational Recommendation," Proc. NAACL-HLT, pp.1490–1504, 2024. doi:10.18653/v1/2024.naacl-long.83.

[12] J. Xu et al., "Qwen3-Omni Technical Report," arXiv:2509.17765, 2025.

[13] Y. Li et al., "MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training," Proc. ICLR, 2024. arXiv:2306.00107.

[14] Qwen Team, "Qwen3.8-27B," model card, 2026. https://huggingface.co/Qwen/Qwen3.8-27B.

[15] D. J. C. MacKay, "The Evidence Framework Applied to Classification Networks," *Neural Computation*, vol. 4, no. 5, pp. 720–736, 1992. doi:10.1162/neco.1992.4.5.720.