Overview and Analysis of the RecSys Challenge 2026: Conversational Music Recommendation

Seungheon Doh
Sony Group Corporation
Japan
seungheon.a.doh@sony.com

Sergio Oramas
SiriusXM
United States
sergio.oramas@siriusxm.com

Bruno Sguerra
Deezer Research
France
bmassonisguerra@deezer.com

Abhinav Bohra
Amazon
United States
bohraabhinav8@gmail.com

Claudio Pomo
Politecnico di Bari
Italy
claudio.pomo@poliba.it

Francesco Barile
Maastricht University
Netherlands
f.barile@maastrichtuniversity.nl

Abstract

The RecSys Challenge 2026 studies conversational music recommendation as a joint item recommendation and response generation problem: given a multi-turn dialogue, systems must retrieve relevant tracks from a large catalog and produce a grounded natural-language response. This paper presents the challenge task, dataset, evaluation protocol, and official results. Beyond the leaderboard, we analyze the 16 accepted systems through a common retrieve–rerank–generate framework and examine how recommendation performance varies across users, requests, and dialogue contexts. Strong systems commonly combine heterogeneous candidate sources and preserve source-specific evidence for learned reranking. Across the system papers and our organizer-side analysis, robust design also means 1) grounding cold-start retrieval in multi-turn conversation and item signals, 2) using intent detectors, and 3) modeling the full multi-turn context rather than the current query alone. We further identify limitations of the benchmark and evaluation protocol, including single-ground-truth relevance and teacher-forced evaluation of synthetic dialogues. Together, these findings provide practical guidance for future conversational recommender systems and shared evaluation efforts.

CCS Concepts

• **Information systems** → **Recommender systems**; **Music retrieval**; • **Human-centered computing** → **Human computer interaction (HCI)**.

Keywords

conversational recommender systems, music recommendation

ACM Reference Format:
Seungheon Doh, Sergio Oramas, Bruno Sguerra, Abhinav Bohra, Claudio Pomo, and Francesco Barile. 2026. Overview and Analysis of the RecSys Challenge 2026: Conversational Music Recommendation. In *RecSys Challenge 2026 (RecSysChallenge '26), October 02, 2026, Minneapolis, MN, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3842413.3842414

This work is licensed under a Creative Commons Attribution 4.0 International License.
RecSysChallenge '26, Minneapolis, MN, USA
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2863-1/2026/10
https://doi.org/10.1145/3842413.3842414

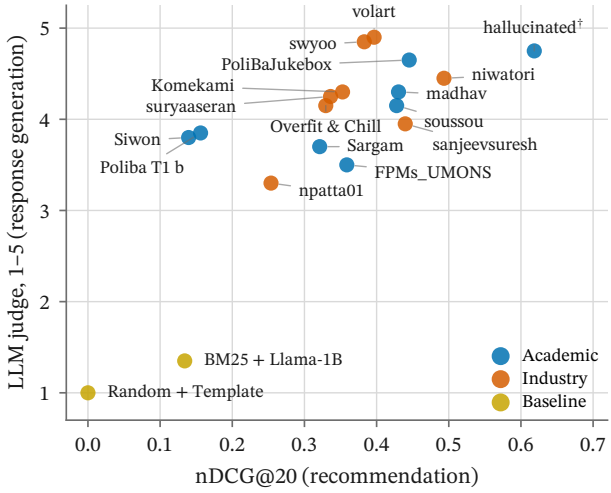

Figure 1: Challenge at a glance: recommendation quality (nDCG@20) versus response quality (LLM-as-a-judge, 1–5). † The evaluation-data-exposed submission.

1 Introduction

The RecSys Challenge 2026 focuses on conversational music recommendation [4, 5, 7, 14, 15]. Conversational recommender systems (CRS) let users express, refine, and receive recommendations through natural, multi-turn dialogue, replacing the static, one-shot lists of traditional recommenders [3, 8, 32]. This is equally critical for music listening: through dialogue, listeners can express their musical preferences, refine them turn by turn, and come to understand the music itself, rather than specifying a fixed preference upfront. A request can identify an exact title, describe an acoustic texture or cover image, ask for a mood or activity, or react implicitly to tracks played earlier, so a useful system must retrieve precisely from a large catalog, update its belief about intent over turns, and explain its choices without inventing evidence. Realizing this requires coupling natural language understanding with high-precision recommendation: at every turn the system must perform two tightly coupled functions, (i) *music recommendation*, ranking relevant tracks from a large catalog given the dialogue history and user context, and (ii) *response generation*, producing

a natural-language reply that justifies the recommendation and sustains the conversation.

This paper first documents the 2026 task, dataset, evaluation protocol, and official results (Sections 2–4). We then go beyond the leaderboard in three ways. Section 5 surveys the 16 accepted systems and their retrieval, reranking, and response-generation components. Section 6 revisits every submission by dialogue depth, warm versus cold condition, goal specificity, and goal topic, using organizer-side resources that were not released to participants (the unmasked source conversations, their goal annotations, and the hidden target tracks). Sections 7 and 8 distill the systems into reusable design principles and audit the benchmark itself, including an evaluation-data exposure in the winning submission and limitations of the dataset and protocol.

Our analysis leads to four main conclusions. (1) Strong systems combine candidates from complementary retrievers and learn from source-specific ranks, scores, and presence together with cross-source agreement. (2) Cold-start pipelines should rely by default on conversation and item signals that remain available without a user ID or profile. (3) Rather than attempting to recover a complete hidden goal taxonomy, systems should use high-precision, abstaining intent detectors to trigger specific ranking actions. (4) Recommendation intent is distributed across turns, so systems should model the full dialogue rather than focus only on the current query. Together, these findings offer an evidence-based blueprint for designing future conversational music recommendation systems.

2 The Challenge Task

Problem Formulation. Let $\mathcal{D} = \{d_1, \ldots, d_N\}$ be the track catalog, where each track d_j carries multimodal content: metadata (title, artist, album, release date), semantic tags, lyrics, audio, and album image. Each user u is associated with a profile p_u of demographic attributes (age group, gender, country). A conversation between the user and the system unfolds over T turns as $s = \{(q_i, m_i, r_i)\}_{i=1}^{T}$, where at each turn i the user issues a natural-language query q_i, and the system responds with a ranked list of tracks $m_i \subset \mathcal{D}$ together with a natural-language response r_i. The task is to model the system at each turn. Concretely, at turn t a system f is given the user profile p_u, the conversation history $s_{t-1} = \{(q_i, m_i, r_i)\}_{i=1}^{t-1}$, and the current query q_t, and must produce both a ranked recommendation list and a response $(m_t, r_t) = f(p_u, s_{t-1}, q_t)$. The two outputs define complementary subtasks.

Task 1. Music Recommendation. The recommendation output m_t is a ranked list of the top 20 tracks from $\mathcal{D}$, which must capture both the explicit request and the implicit preferences expressed across the conversation history s_{t-1}, the current query q_t, and musical taste of user u.

Task 2. Response Generation. The response output r_t is a natural-language reply that justifies the recommended tracks m_t and maintains a coherent conversational flow with the user, conditioned on (p_u, s_{t-1}, q_t).

Participation. The challenge ran on Codabench [1] from April 10 to June 30, 2026, with separate academic and industry tracks and two evaluation phases, interim (Blind-A) and final (Blind-B). It attracted 336 registered teams and 9,554 submissions, of which 40 teams reached the Blind-B leaderboard; 20 teams submitted system papers, of which 16 were accepted after review. The organizers also released two reference baselines: a random recommender with template responses, and BM25 retrieval over the dialogue paired with a Llama-1B response generator.

[1] https://www.codabench.org/competitions/15786

Table 1: Statistics of the TalkPlayData-Challenge splits. #Tracks is the number of distinct tracks. #Turns is the mean turns per conversation; $|q|$ and $|r|$ are mean query and response lengths in words.

Split	#Conv.	#Users	#Tracks	#Turns	$\lvert q\rvert$	$\lvert r\rvert$
Training Set	15,199	8,591	43,597	8.0	44.7	36.6
Dev Set	1,000	500	6,761	8.0	40.8	44.6
Blind-A Set	80	58	208	3.6	39.0	33.5
Blind-B Set	80	54	274	4.5	42.1	35.2

3 Dataset

We source conversation sessions from the *TalkPlayData-Challenge* dataset [2], a synthetic corpus constructed via a multimodal agentic pipeline grounded in the LFM-2b dataset [20], in which a Listener LLM and a Recommender LLM engage in a goal-conditioned dialogue, yielding sessions of up to eight turns grounded in real user demographics, listening histories, and track metadata. The data is split into three parts: participants train and validate on a public release[2] of 15,199 training and 1,000 development conversations, and are then ranked on two held-out blind sets of 80 sessions each, *Blind-A*[3] and *Blind-B*[4]. Ground-truth responses and target tracks are never disclosed for the blind sets, which are drawn from distinct conversations to measure generalization rather than overfitting: Blind-A drives the live leaderboard, while Blind-B, released near the deadline, determines the final ranking. Blind-B additionally masks the user identifier and profile in 40 of its 80 sessions; for analysis we call the profile-visible half *warm* and the masked half *cold*, and Section 8 explains why this is an operational missing-profile condition rather than a fully disjoint-user cold-start test.

To let participants focus on modeling rather than feature extraction, we provide four supporting resources over the shared catalog of 47,071 tracks and 8,772 users. *Track Metadata*[5] and *User Metadata*[6] provide the official track catalog (title, artist, album, tags, etc.) and the demographic attributes of the user profile p_u. We further release pre-extracted *Track Embeddings*[7] spanning six modalities: audio (CLAP [30]), album image (SigLIP-2 [26]), text-based tags, lyrics, and metadata (Qwen3-Embedding 0.6B [33]), and a collaborative-filtering embedding (BPR [19]). Matching *User Embeddings*[8] are also provided for the collaborative-filtering side.

[2] https://huggingface.co/datasets/talkpl-ai/TalkPlayData-Challenge-Dataset
[3] https://huggingface.co/datasets/talkpl-ai/TalkPlayData-Challenge-Blind-A
[4] https://huggingface.co/datasets/talkpl-ai/TalkPlayData-Challenge-Blind-B
[5] https://huggingface.co/datasets/talkpl-ai/TalkPlayData-Challenge-Track-Metadata
[6] https://huggingface.co/datasets/talkpl-ai/TalkPlayData-Challenge-User-Metadata
[7] https://huggingface.co/datasets/talkpl-ai/TalkPlayData-Challenge-Track-Embeddings
[8] https://huggingface.co/datasets/talkpl-ai/TalkPlayData-Challenge-User-Embeddings

Table 2: Official Blind-B results for all teams with an accepted system paper. Recommendation diversity is catalog coverage, lexical diversity is Distinct-2, and the LLM-as-a-judge score is on its original 1–5 scale. † The evaluation-data-exposed submission.

				Recommendation System		Response Generation	
Rank	Team	Track	Final Score	nDCG@20	C. Diversity	LLM Judge	L. Diversity
1	hallucinated [1]†	Academic	0.689	0.618	0.027	4.750	0.957
2	volart [28]	Industry	0.587	0.397	0.032	4.900	0.927
3	niwatori [29]	Industry	0.586	0.493	0.031	4.450	0.774
4	swyoo [31]	Industry	0.578	0.383	0.030	4.850	0.952
5	PoliBaJukebox [12]	Academic	0.576	0.445	0.030	4.650	0.765
9	madhav [16]	Academic	0.548	0.430	0.031	4.300	0.817
11	soussou [22]	Academic	0.530	0.428	0.031	4.150	0.771
12	sanjeevsuresh [24]	Industry	0.528	0.440	0.032	3.950	0.833
16	Komekami [25]	Industry	0.500	0.353	0.028	4.300	0.731
17	suryaaseran [21]	Industry	0.492	0.336	0.029	4.250	0.773
20	Overfit & Chill [23]	Industry	0.477	0.330	0.031	4.150	0.733
24	FPMs_UMONS [13]	Academic	0.451	0.359	0.031	3.500	0.810
25	Sargam [18]	Academic	0.448	0.321	0.012	3.700	0.840
29	npatta01 [17]	Industry	0.381	0.254	0.031	3.300	0.786
31	Poliba T1 b [27]	Academic	0.366	0.156	0.031	3.850	0.710
34	Siwon [10]	Academic	0.354	0.140	0.031	3.800	0.713
–	BM25 + Llama-3.2-1B-Instruct	Baseline	0.162	0.134	0.020	1.350	0.664
–	Random + Template Response	Baseline	0.067	0.000	0.033	1.000	0.632

Each conversation goal has two axes [2]. The topic code covers audio (A, e.g. "discover songs with immersive soundscapes"), lyrics (B), visual–musical connections (C), contextual or situational use (D), interactive refinement (E), metadata-rich exploration (F), mood and emotion (G, e.g. "I need something to cheer me up"), artist/discography (H), culture/geography (I), social/popularity context (J), and temporal/era discovery (K, e.g. "music from the 80s please"). Specificity is a two-letter code: the first letter is query specificity and the second target specificity. LL (e.g. "play some chill music") and HL (e.g. "find bebop jazz with saxophone, 1950s–60s") admit many relevant tracks; LH (e.g. "what was the popular song from a recent musical movie?") describes a vague clue with one or few intended targets; and HH (e.g. "Wonderwall by Oasis") is a precise lookup.

4 Evaluation

The official score combines what the system recommends and how it describes the recommendation: Score = $0.50\,\text{nDCG@20} + 0.10\,\text{Catalog Diversity} + 0.10\,\text{Lexical Diversity} + 0.30\,\text{LLMJudge}_{\text{norm}}$.

For each evaluated session-turn e, only one track is labeled relevant. If it appears at rank $k \leq 20$, its nDCG contribution is $1/\log_2(k+1)$, and otherwise zero. The official aggregate first averages within each target-turn index and then across turn indices. Blind-B deliberately contains ten targets at each of turns 1–8, so its official macro average equals the ordinary 80-session mean.

Catalog diversity is the fraction of catalog tracks that appear at least once across all submitted lists. Lexical diversity is corpus-level Distinct-2 [11]. On a fixed subset of 10 Blind-B sessions, the LLM judge [6, 34] scores personalization and explanation quality from 1 to 5; the two dimension means are averaged and normalized as $(\text{Judge} - 1)/4$ for the score above. The final evaluator used Gemini-3.1-Flash-Lite [9].

Official results. Table 2 groups the component metrics by subtask. The winner, hallucinated, leads ranking by a wide margin (0.618 nDCG@20 versus 0.493 for the next system). Although the team did not use Blind-B's hidden ground-truth answers, its paper discloses that the Blind-B dataset itself was used for training: fusion weights were tuned and the final reranker trained on its conversation turns. We flag this evaluation-data exposure by shading throughout this paper (Section 8). The two subtasks are only loosely coupled: the best judge scores belong to volart (4.90) and swyoo (4.85), whose rankings are mid-pack, and rank-2 volart trails rank-3 niwatori by 0.10 nDCG while beating it by 0.45 judge points. Every accepted system beats both organizer baselines on ranking, although the margin is small for the lowest-ranked entries. Catalog diversity is nearly constant (0.03) because 80 sessions × 20 slots cap it at 0.034 of the 47,071-track catalog, so it does not discriminate between systems [28].

5 Submitted Approaches

The 16 accepted papers span systems and diagnostics, but converge on a modular retrieve–rerank–generate cascade. Table 3 summarizes each team's retrieval, reranking, and response-generation components. Among the non-exposed systems, runner-up volart pairs a low-cost four-lane retrieval pool with LambdaMART and critic-guided response editing, arguing that ranking is information-limited once the candidate neighborhood is found, and posts the highest LLM-judge score of any submission [28]. Third-place niwatori uses 14 retrieval sources and shows that retaining each source's rank, score, and presence is more effective than collapsing the union too early, giving it the strongest or joint-strongest ranking results among non-exposed systems on several goal topics (Section 6) [29]. swyoo's two-view retrieval and claim-validated generation pipeline

Table 3: Component summary of the 16 accepted systems, ordered by final rank: retrieval (candidate generation), reranking/fusion, and response generation. Multi-component pipelines are condensed to their load-bearing pieces. † The evaluation-data-exposed submission.

Team	Retriever (Candidate Generation)	Reranker	Response generator
hallucinated [1]†	BM25, dense, hybrid two-tower, ItemKNN, six multimodal sequential generators, all CVaR-tuned	Weighted RRF; XGBoost LTR ensemble; continuity/query heuristics	Gemma-4-26B-A4B-it, summarize–generate–compress–diversify
volart [28]	BM25, OpenAI text-embedding-3-large, LLM track-description, co-occurrence	RRF pool; LambdaMART with missing-field-safe features	Best-of-3 Gemini-3.1-Flash-Lite; independent gpt-4o-mini critic edit
niwatori [29]	14 lexical, dense, entity, history, transition, co-occurrence sources	LightGBM LambdaRank with source features	Qwen3.6-27B grounded on top-3; diversity selection
swyoo [31]	BM25 + frozen Qwen3-Embedding-0.6B hybrid pool; QLoRA-tuned Qwen3-Embedding-8B conversational pool	Top-100 union; LightGBM with cross-pool RRF and request-regime features	GLM-5.2 propose–assign–select with deterministic claim validation
PoliBaJukebox [12]	10 lexical, collaborative, CLAP acoustic, organizer Qwen3 embeddings and fine-tuned BGE-base/BGE-large sources	Weighted RRF; LightGBM LambdaRank + CatBoost YetiRank; density-ratio weighting	Two-pass fact-grounded generation (Gemini-3.1-Pro-Preview draft, Gemini-3.5-Flash critic)
madhav [16]	Fine-tuned MPNet dense retrieval with continuity boosts	Rank-zone fence (popularity/co-occurrence); LightGBM tail reranker	Gemini-2.5-Pro narrates only the clean top-3 head
soussou [22]	Broad, feedback-aware, and fielded BM25; fine-tuned BGE-base-en-v1.5; structural; visual	LightGBM LambdaRank with source features	Qwen3-14B multi-candidate; heuristic selection; constrained rewrite
sanjeevsuresh [24]	BM25, dense text (metadata/lyrics/tags), query-expansion and album-artwork arms	XGBoost LambdaRank (49 features); exact-request detector gates a specialist trained with request-satisfying targets	LLM responder (model N/A)
Komekami [25]	8 lexical anchor + 12 continuity/neighbor/entity/transition sources	Anchor LightGBM; LightGBM LambdaRank on 73 rank-only features	Claude Opus 4.8 on top-1 metadata; ranking frozen
suryaaseran [21]	BM25, LoRA-tuned multilingual-e5-base and frozen Qwen3-Embedding-0.6B, history-neighbor, artist expansion, CF-BPR, co-occurrence, SASRec	LightGBM LambdaMART with 67 features	Decoupled responder (model N/A)
Overfit & Chill [23]	Qwen3-8B query planner; 10 text, anchor, identity, training-mined channels	LightGBM LambdaRank with source-aware features	Same Qwen3-8B with three state-routed templates
FPMs_UMONS [13]	RRF over BM25, fine-tuned Qwen3-Embedding-4B, item-CF, artist expansion	Qwen3-8B listwise scoring head over top-200	Best-of-20 Gemma-3n-E4B selected by a Gemma-4-E2B judge.
Sargam [18]	Artist/album resolver; BM25; multi-seed CF/audio/visual and user-CF neighbors	Weighted RRF with exact-target override	Constrained Gemini response on the top-1 track (version N/A)
npatta01 [17]	11 BM25, dense-text, CLAP text-to-audio, anchor-centroid, user-CF, lookup branches	LambdaMART over branch, bi-encoder cosine and typed-state features	Qwen3-30B-A3B-Instruct-2507 conditioned on typed state
Poliba T1 b [27]	Heterogeneous user–session–track GNN; BM25; CLAP audio; Qwen3 lyric embeddings	Query-dependent neural router + weighted RRF; LLM rerank + reflective replacement	Reasoning draft + persona/style rewrite (model N/A)
Siwon [10]	BM25; frozen Qwen3-Embedding-4B; listening-history centroid	RRF	Frozen Qwen3-32B (thinking) rationale, frozen Phi-4 summarizer

†The evaluation-data-exposed submission (Section 8).

places fourth with the second-highest judge score of any submission [31], and PoliBaJukebox combines ten retrieval sources with adversarial-validation-based correction for dev-to-blind deployment shift [12]. The top-ranked submission, hallucinated, combines broad lexical, dense, two-tower, nearest-neighbor, and sequential candidate generators, each tuned under a CVaR objective, with weighted RRF fusion and an XGBoost ranker [1]; however, its paper discloses that fusion weights and reranker were tuned and trained on Blind-B conversation turns, so we treat its numbers as descriptive only and set its results in gray throughout (Section 8).

5.1 Retriever and Reranker

Nearly every strong submission uses heterogeneous candidate sources. Common components are fielded BM25 and exact entity matching, dense query–track retrieval, artist/album expansion, item or session co-occurrence, and collaborative signals. Sargam adds multi-seed CF, audio, and visual neighbors and a high-specificity catalog resolver [18]; FPMs_UMONS combines BM25, dense retrieval, item-CF, and artist expansion [13]; Poliba T1 b's graph-aware entry joins a heterogeneous user–session–track encoder with sparse, audio, and lyric retrievers under a learned router [27].

The dominant fusion pattern is a high-recall union followed by a tree-based learning-to-rank model. niwatori, volart, PoliBaJukebox, swyoo, suryaaseran, Komekami, soussou, Overfit & Chill, npatta01, and hallucinated all preserve source-specific evidence for LightGBM or XGBoost reranking [1, 12, 17, 21–23, 25, 28, 29, 31]. FPMs_UMONS instead trains an LLM scoring head and shows that selecting the top item and ordering ranks 2–20 are distinct skills [13]. Weighted RRF remains a strong training-free fallback and candidate-pool constructor [10, 18]. Two regime-aware ideas are particularly reusable. PoliBaJukebox detects dev-to-blind covariate shift by adversarial validation and trains its rerankers with tempered density-ratio importance weights, a training-time correction that leaves inference unchanged [12]. madhav's Compass fences noisier popularity/co-occurrence signals into ranks 4–20 so that the generated response describes only a clean head, and its

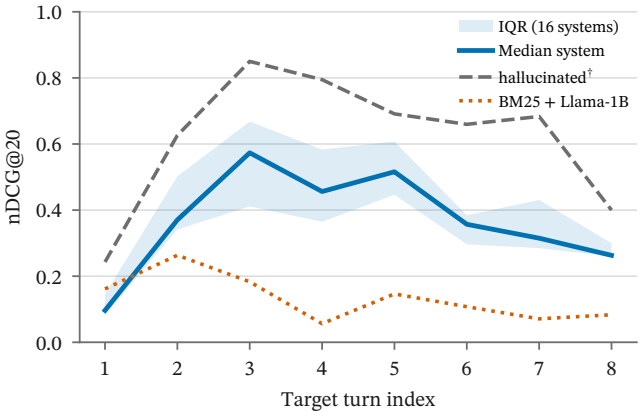

Figure 2: Turn-macro nDCG@20 by target turn index over Blind-B. † The evaluation-data-exposed submission.

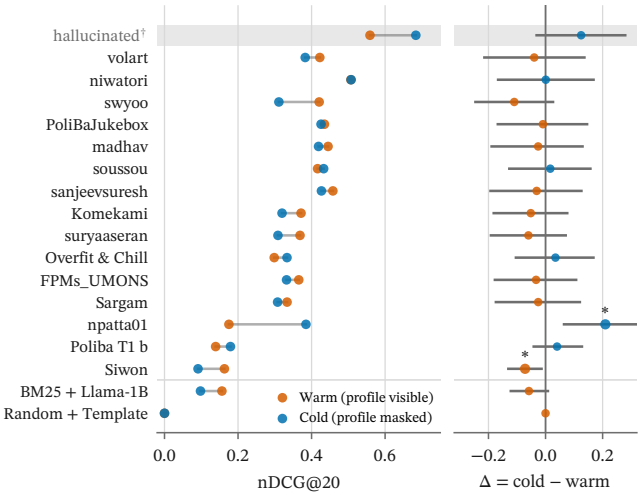

Figure 3: Warm versus cold nDCG@20 per system (left) and the cold-minus-warm contrast with stratified-bootstrap 95% intervals (right). * interval excludes zero. † The evaluation-data-exposed submission.

analysis shows that gating continuity by predicted continuation-versus-discovery intent recovers the novel-artist slice [16]. These methods adapt to deployment regime and intent rather than assuming one fusion rule is universally correct.

5.2 Grounded Response Generation

Most teams decouple ranking from prose: once top-20 tracks are fixed, the generator receives visible dialogue and verified metadata. swyoo's propose–assign–select pipeline groups candidate evidence before writing and validates title and attribution claims deterministically [31]. volart and FPMs_UMONS generate several responses, then apply independent critics or deterministic filters for grounding and diversity [13, 28]. soussou uses constrained rewrites only when a heuristic quality score improves [22]. The common lesson is that the language model may describe catalog facts and conversational evidence but must not silently change the ranked decision.

6 Recommendation Performance Analysis

We join four organizer-side resources by (session_id, turn): public Blind-B inputs, unmasked source conversations with goal annotations, the hidden target file, and each submitted prediction. A session is cold-start exactly when the public Blind-B row masks its user ID/profile. For every system and slice, we compute nDCG@20 per session, average within each represented target turn, and then average across turns, matching the official macro convention. Warm/cold uncertainty uses 10,000 bootstrap resamples within each condition–turn cell. Goal-topic results are descriptive because topics contain only 2–12 sessions.

6.1 Turn Depth

Figure 2 suggests that longer dialogue contexts remain difficult for current systems: median nDCG rises from 0.10 at turn 1 to 0.57 at turn 3, but falls to 0.26 by turn 8. Turn 1 offers no played-track history, so continuation and co-occurrence sources are silent and systems depend entirely on the first request. The later decline matches Siwon's development finding that nDCG@20 falls monotonically with turn index and that request-focused query construction does not recover it [10], and motivates the dual-view retrieval (a literal view of the latest request beside a full-dialogue view) used by several teams [22, 31]. The BM25 baseline stays low at every depth, which is consistent with history-aware sources contributing to the mid-conversation gains, although this comparison does not isolate their causal effect.

6.2 Warm Start versus Cold Start

Figure 3 compares nDCG@20 across 40 warm sessions with visible user IDs and profiles and 40 cold sessions in which both are masked. It reveals three regimes. First, hallucinated's cold score (0.684) exceeds its warm score (0.559) by the largest margin of any system. Before its disclosure this would read as remarkable cold-start robustness from query/session evidence, worst-slice CVaR tuning, and continuity heuristics [1]. Given that its candidate fusion and reranker were tuned and trained on Blind-B conversation turns, the asymmetry cannot be attributed to architecture alone.

Second, the only non-exposed contrast whose interval excludes zero on the cold side belongs to npatta01 (+0.210): its typed-state, text-first retrieval treats the profile as one optional branch [17], although the asymmetry equally reflects a weak warm-side score (0.175), so it is evidence of profile independence rather than of overall strength. More broadly, niwatori is essentially invariant (0.507 versus 0.508), while soussou and Overfit & Chill are modestly higher on cold. Their shared property is that identity is not foundational: niwatori and soussou use no profile or user-identity signal at all, drawing history only from tracks observed in the dialogue [22, 29], and Overfit & Chill's user-to-item BPR channel is one of ten in its union [23].

Third, most systems favor warm sessions, but only Siwon's −0.072 contrast excludes zero in that direction [10]. Among point estimates, swyoo shows the largest warm advantage (−0.110), whereas PoliBaJukebox is nearly invariant despite using no profile features and reweighting training toward blind-like sessions [12].

	A Audio n=4	B Lyrics n=11	C Visual n=9	D Context n=9	E Refine n=8	F Metadata n=3	G Mood n=11	H Artist n=12	I Culture n=2	J Popular n=7	K Era n=4
hallucinated†	.44	.55	.58	.75	.69	.78	.68	.38	1.0	.64	.58
volart	.50	.40	.45	.41	.45	.67	.44	.32	.50	.44	.25
niwatori	.59	.57	.48	.54	.72	.67	.56	.27	.50	.40	.25
swyoo	.59	.39	.53	.48	.41	.54	.43	.26	.50	.22	.13
PoliBaJukebox	.42	.51	.56	.46	.43	.67	.43	.26	.50	.36	.50
madhav	.50	.46	.45	.40	.65	.67	.47	.22	.50	.38	.31
soussou	.53	.31	.67	.47	.59	.67	.47	.19	.50	.31	.16
sanjeevsuresh	.52	.51	.51	.46	.62	.67	.50	.23	.50	.31	.16
Komekami	.39	.35	.46	.33	.49	.54	.35	.22	.50	.32	.16
suryaaseran	.37	.38	.46	.46	.45	.50	.41	.17	.50	.18	.00
Overfit & Chill	.32	.52	.61	.38	.47	.50	.25	.15	.17	.34	.06
FPMs_UMONS	.41	.36	.48	.38	.46	.54	.40	.22	.22	.31	.12
Sargam	.49	.34	.46	.31	.40	.54	.29	.10	.50	.32	.25
npatta01	.33	.40	.27	.30	.11	.42	.23	.14	.19	.42	.12
Poliba T1 b	.19	.18	.14	.12	.17	.14	.11	.11	.00	.22	.22
Siwon	.18	.16	.20	.15	.07	.12	.14	.10	.13	.17	.06
BM25 + Llama-1B	.25	.13	.16	.09	.02	.17	.11	.13	.19	.22	.21
Random + Template	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00

Figure 4: Turn-macro nDCG@20 by conversation-goal topic. † The evaluation-data-exposed submission.

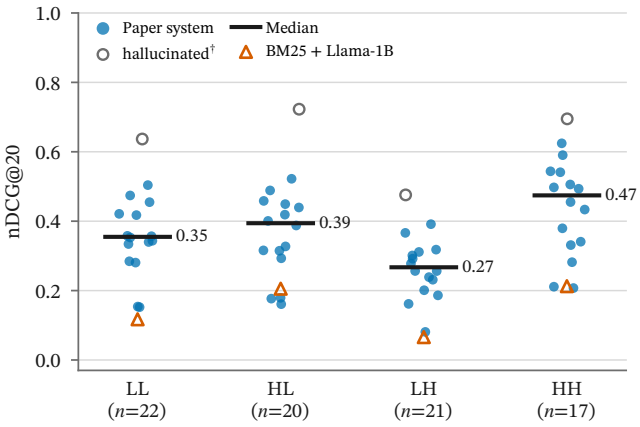

Figure 5: Per-system nDCG@20 by goal-specificity regime. Precise lookups (HH) are easiest and vague single-target requests (LH) hardest. † The evaluation-data-exposed submission.

The profile-agnostic BM25 baseline shows a similar warm advantage (−0.058; 0.098 cold versus 0.156 warm), suggesting that the masked half is intrinsically harder. Thus, the dominant pattern is a cold-session penalty, although it cannot be attributed solely to profile removal. Wide intervals across the 40-session halves preclude a definitive cold-start ranking.

6.3 Goal Specificity

Across the 16 systems, median nDCG is 0.355 on LL, 0.394 on HL, 0.267 on LH, and 0.474 on HH (Figure 5). HH benefits from explicit titles, artists, albums, and other catalog-resolvable constraints; nearly every pipeline contains exact matching or entity expansion. In contrast, LH asks the system to recover one or a few intended tracks from vague clues. This combines low observable information with a single-target metric, matching the information-limit diagnosis that dialogue identifies a neighborhood more reliably than the logged point within it [28]. The same LH dip appears in the BM25 baseline (0.066 versus 0.213 on HH), so the regime ordering is a property of the task, not of any modeling family.

The gap also explains why an exact-target module is best treated as a routed specialist rather than a universal semantic reranker. High-precision catalog resolution can help HH [18, 24], while LH needs diversified recall, uncertainty-aware ranking, or clarification. The exposed hallucinated submission posts the highest value on all four slices; among non-exposed systems, niwatori leads LL, LH, and HH while volart leads HL. Thus no one modality explains performance across the specificity axis.

6.4 Goal Topic

Figure 4 summarizes topic difficulty across every submission. Artist/discography (H, median 0.219) and temporal/era (K, 0.158) are hardest, followed by social/popularity context (J, 0.319). The artist result is not evidence that entity matching fails; rather, many H goals ask for discovery within or beyond a discography, where the logged continuation policy, artist repetition, and user request may conflict [16, 25]. Era queries combine fuzzy temporal language with incomplete or noisy release metadata.

Visual (C, median 0.472) and metadata-rich (F, 0.544) appear easier, with soussou strongest on C, consistent with its explicit last-item and history-image retrievers [22]. These slices contain nine and three sessions, respectively, so this is a hypothesis for a larger test, not an ablation. The same caveat applies to culture/geography (I): the exposed hallucinated submission scores a perfect 1.0, but $n = 2$. Audio (A) is led jointly by niwatori and swyoo (0.589 each), and lyrics (B) by niwatori, whose broad source union covers both lexical and semantic evidence [29].

7 Reusable Design Principles

The system papers and our stratified analysis yield four evidence-based design principles for future conversational music recommenders. We ground each principle in reported system components or controlled ablations, distinguish submitted methods from analysis-only findings, and use the organizer-side slices in Section 6 as supporting rather than causal evidence.

1. Combine candidates from multiple retrievers, then learn which sources to trust. Use lexical, dense, entity, history, co-occurrence, and optional identity or multimodal sources as parallel candidate generators: each fails in a different regime, so their union sets a high recall ceiling cheaply. The reranker should then use source-specific evidence: which retrievers proposed a track, at what rank, with what score, and how many retrievers agree. Ten of the 16 accepted systems pass such signals to a tree-based learning-to-rank model [1, 12, 17, 21–23, 25, 28, 29, 31]. niwatori provides a controlled comparison over the same candidate union: RRF reaches 0.1715 nDCG@20, a LambdaRank model using only context–track features reaches 0.1832, and adding per-retriever rank, score, presence, and agreement features raises it to 0.1994 [29]. When no training budget exists, weighted RRF over the same union remains a strong fallback and candidate-pool constructor [10, 18].

2. Use conversation and item signals to handle cold-start. Cold-start robustness comes from grounding the ranking pipeline in information that remains observable without a user profile. niwatori does not use user IDs, profiles, or user embeddings for ranking. Its BM25, TF–IDF, tag/entity, and two-tower sources retrieve from the current message and observed dialogue, while artist/album expansion and co-occurrence/transition sources use tracks already observed within the session [29]. npatta01 illustrates a similar fallback architecture: it compiles the dialogue into a typed conversation state and makes user-CF only one of eleven retrieval branches, so text, entity, and item-based branches can operate without it [17]. This is architectural rather than performance evidence, since its paper reports that an in-sample estimate of 0.3844 nDCG@20 fell to 0.2032 under out-of-fold evaluation. Stronger validation comes from volart, which masks the goal and half of the profiles on held-out data and finds essentially no loss after the corresponding features are nulled (goal $\Delta = -0.0025$, $p = 0.31$) [28]. The converse case is Overfit & Chill: two unavailable goal fields collapsed to an unseen sentinel value on Blind-B, and retraining without them raised the composite from 0.42 to 0.48 [23]. Together, these results favor conversation- and item-based retrieval as the default path, with identity, profiles, and collaborative signals treated as optional additions.

3. Use targeted intent detectors for specific ranking actions. The organizer-side goal categories and HH/HL/LH/LL specificity labels used in our analysis were not visible in Blind-B, so the evidence supports detecting narrow, actionable intents rather than recovering the full hidden taxonomy. sanjeevsuresh detects exact-song and version requests with regular expressions over quoted titles and imperative phrases, resolves them against title and artist fields in the catalog, and abstains when the match is ambiguous. The detector gates a specialist reranker trained with request-satisfying targets in addition to the official labels, raising nDCG@20 from 0.523 to 0.802 on the exact/version conflict slice without reducing the whole-split official metric [24]. For continuation versus discovery, Compass trains a logistic classifier on the dialogue-query embedding to predict whether the target artist already appears in the session history. Its predicted probability scales the artist/album continuity boost, reaching AUC 0.788 and improving the novel-artist slice in local analysis; however, this gate was analysis-only and did not yield an overall Blind-B gain [16]. Komekami likewise uses high-precision lexical rules for phrases such as "a different artist" or "something new," but only to audit artist-change intent rather than route its submitted ranker [25]. These results favor high-precision detectors that abstain under ambiguity and trigger a defined ranking action; they do not establish general goal or specificity prediction.

4. Do not focus only on the current query; consider the multi-turn context. Recommendation intent is distributed across the dialogue: the current request may refine, reject, or refer back to preferences and tracks introduced in earlier turns. Systems should therefore model the conversation as a whole rather than reduce it to the latest utterance. In Siwon's controlled dense-retrieval comparison, the full-dialogue query reaches 0.091 nDCG@20, while three request-focused variants score 0.089, 0.085, and 0.082; full-dialogue BM25 is stronger still at 0.118 [10]. swyoo's fusion of a conversational pool with its hybrid pool raises nDCG@20 from 0.2165 to 0.2198 [31]. soussou provides a complementary ablation: removing its contextual dense retriever lowers nDCG@20 from 0.1851 to 0.1778, while removing its recent-utterance dense retriever lowers it to 0.1825 [22]. Our turn-depth analysis is consistent with this principle because later turns remain harder, but it does not by itself identify inadequate context modeling as the cause.

8 Limitations and Future Challenges

Evaluation-data exposure. The winning submission's paper discloses that its RRF fusion weights were tuned to maximize recall on held-out Blind-B target turns and that its XGBoost reranker was model-selected on the last available turn of each Blind-B session and, for the final submission, trained on "all available Blind-B turns except those used for validation" [1]. Although the hidden target tracks themselves were never revealed, the conversational history of every evaluation session, including the sessions whose profiles were masked, entered training. This is outside the intended scenario, in which Blind-B measures generalization to unseen conversations, and it plausibly inflates both the overall score and the cold-start result in Section 6. We keep the official ranking intact, because the challenge rules did not explicitly forbid transductive use of the released blind inputs, but we gray out the affected rows in every table and figure and exclude the system when naming slice leaders. Future editions should state an explicit data-usage contract (no training, tuning, or model selection on blind-phase inputs), require a data-usage declaration at submission time, and audit winning pipelines against it before finalizing awards.

Dataset integrity: Reverse-engineering risk. Because the challenge dataset (TalkPlayData-Challenge) is synthetically generated from LFM-2b listening histories, the benchmark may retain track sequence correspondences to its source data. During the challenge,

we received a report that similarities between multi-turn conversational patterns and listening histories could be exploited to partially reconstruct the latent recommendation pool. Although this does not directly reveal the hidden target, it can reduce the effective search space and highlights the importance of dataset integrity in challenge benchmarks, particularly as synthetic datasets derived from existing corpora become more common.

To address this concern, the organizers conducted additional verification of the award-winning submissions, including not only inference reproduction but also independent reconstruction of the submitted systems from their code and configurations. This process provided assurance that the reported performance could be reproduced without relying on reverse engineering of the challenge dataset. Future editions should explicitly audit source-data linkages and reconstructability as part of benchmark design and validation. We thank the members of team hallucinated [1] for identifying and reporting this issue.

Dataset limitation: Single-ground-truth scenario. Each turn has one target selected by the synthetic recommender from a latent session pool, so open-ended requests may admit many equally useful tracks that the metric counts as failures, and exact requests can conflict with the logged choice (Section 7); npatta01 further reports that two LLM judges relabeling training turns disagree roughly 20% of the time [17]. The one-label problem is most acute in the LH regime, the hardest slice for nearly every system including the BM25 baseline: when a vague query is paired with a single logged target, the metric cannot separate a wrong recommendation from a valid alternative that the policy did not log. In such cases a real assistant would likely ask a clarifying question rather than commit to a list. Future releases should combine the logged action with catalog-resolved constraints and graded, multi-label relevance, report target recovery and request satisfaction separately, and permit an explicit clarification action scored by the information it recovers.

Evaluation limitation: Teacher-forced, synthetic multi-turn evaluation. Every target turn is evaluated against the logged history: systems condition on the synthetic policy's past recommendations, never on their own, so error compounding, recovery from a bad suggestion, and long-horizon preference tracking are untested. The benchmark ultimately measures recovery of a synthetic policy's next track rather than the long-term utility of an interactive recommender. A next edition should complement teacher-forced snapshots with a reproducible simulator or a controlled user study in which systems may ask clarifying questions and receive feedback, scored by turns-to-success, constraint satisfaction, and user effort.

8.1 Encountered Issues

The main issue we faced concerned the scope of the recommendation catalog. The metadata released to participants contained two splits, `all_tracks` and `test_tracks`, with the second one originally included to experiment with the appropriate scope of tracks for the evaluation. Submissions were expected to rank tracks from the full `all_tracks` catalog. This requirement was stated in the baseline GitHub repository from the beginning of the challenge. It was not added to the Codabench description until the start of Blind-B, resulting in inconsistent guidance across the two official sources.

Restricting the recommendations to `test_tracks` drops the candidate pool to a subset much closer to the evaluation data, which inflates the ranking performance. The inconsistency was identified near the end of the competition after being reported by a participating team. Teams affected by the issue were allowed to resubmit using the full catalog, and the `test_tracks` split was removed from the Hugging Face repository to prevent further confusion. All results reported in Table 2 correspond to submissions evaluated over the complete `all_tracks` catalog.

9 Conclusion

The RecSys Challenge 2026 established a shared benchmark for conversational music recommendation, bringing together conversational understanding, large-catalog retrieval, learned ranking, and grounded response generation within a single evaluation setting. Across 16 accepted systems, the clearest retrieval pattern is to combine complementary candidate sources and preserve source-specific ranks, scores, and presence together with cross-source agreement for learned reranking. For cold-start users, the more robust default is to retrieve from the conversation and observable item evidence while treating identity, profiles, and collaborative signals as optional additions.

Our organizer-side analysis also shows that one recommendation policy does not fit every request. The available evidence favors narrow, high-precision intent detectors that abstain under ambiguity and trigger a defined ranking action, rather than attempting to reconstruct a complete hidden goal taxonomy. It also favors modeling the full multi-turn dialogue: current requests often refine, reject, or refer back to evidence from earlier turns, and later turns remain difficult even for strong systems. Vague single-target requests and artist or era discovery remain especially challenging, underscoring the need to represent uncertainty rather than force every conversation into one confident target.

Finally, the challenge exposed important limitations of the benchmark itself, including evaluation-data exposure, single-ground-truth relevance, and teacher-forced synthetic multi-turn evaluation. Future benchmarks should adopt clearer data-usage protocols, a single authoritative task specification kept in sync across all channels, richer and potentially graded relevance judgments, stronger dataset-integrity checks, and genuinely interactive evaluation. Moving beyond recovery of a logged target toward measuring request satisfaction and conversational utility will be essential for evaluating whether future systems provide recommendations that are both relevant and useful.

Acknowledgments

We thank everyone who contributed to organizing the RecSys Challenge 2026, as well as all participating teams for their engagement and contributions. We are also deeply grateful to Sony Group Corporation, SiriusXM, and Deezer for their generous support of the challenge prizes, which helped recognize the outstanding efforts of the participating teams.

References

[1] Abdallah Alkhetiar, Luigi Inguaggiato, Nicolò Locatelli, Roberto Manea, Alessio Pizzini, David Ravelli, Gianmarco Schifone, Matteo Vitali, Andrea Zhang, Michael Benigni, Andrea Pisani, and Maurizio Ferrari Dacrema. 2026. Talk Less, Rank Better: A Decoupled Pipeline for Conversational Music Recommendation. In *Proceedings of the RecSys Challenge Workshop at the 20th ACM Conference on Recommender Systems (RecSys Challenge '26)*. ACM, Minneapolis, MN, USA. doi:10.1145/3842413.3842429

[2] Keunwoo Choi, Seungheon Doh, and Juhan Nam. 2025. TalkPlayData 2: An Agentic Synthetic Data Pipeline for Multimodal Conversational Music Recommendation. *CoRR* abs/2509.09685 (2025). doi:10.48550/arXiv.2509.09685

[3] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards Conversational Recommender Systems. In *KDD*. ACM, 815–824.

[4] Seungheon Doh, Keunwoo Choi, and Juhan Nam. 2025. TalkPlay: Multimodal Music Recommendation with Large Language Models. *CoRR* abs/2502.13713 (2025). doi:10.48550/arXiv.2502.13713

[5] Seungheon Doh, Keunwoo Choi, and Juhan Nam. 2025. TalkPlay-Tools: Conversational Music Recommendation with LLM Tool Calling. *CoRR* abs/2510.01698 (2025). doi:10.48550/arXiv.2510.01698

[6] Seungheon Doh, Bruno Sguerra, Sergio Oramas, Elena V. Epure, and Juhan Nam. 2026. LLM-as-a-Judge for Evaluating System Responses in Conversational Music Recommendation. In *Proceedings of the 20th ACM Conference on Recommender Systems* (Minneapolis, MN, USA) *(RecSys '26)*. Association for Computing Machinery, New York, NY, USA.

[7] Elena V. Epure, Yashar Deldjoo, Bruno Sguerra, Markus Schedl, and Manuel Moussallam. 2025. Music Recommendation with Large Language Models: Challenges, Opportunities, and Evaluation. *CoRR* abs/2511.16478 (2025). Under review, ACM Transactions on Recommender Systems. doi:10.48550/arXiv.2511.16478

[8] Mehmet H. Göker and Cynthia A. Thompson. 2000. The Adaptive Place Advisor: A Conversational Recommendation System. In *Proceedings of the 8th German Workshop on Case Based Reasoning*. 187–198.

[9] Google DeepMind. 2026. Gemini 3.1 Flash-Lite Model Card. https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/. Accessed: 2026-08-20.

[10] Siwon Lee and Young-June Choi. 2026. An Empirical Analysis of Retrieval and Evaluation in the RecSys Challenge 2026 Music-CRS Track. In *Proceedings of the RecSys Challenge Workshop at the 20th ACM Conference on Recommender Systems (RecSys Challenge '26)*. ACM, Minneapolis, MN, USA. doi:10.1145/3842413.3842422

[11] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *NAACL-HLT*. ACL.

[12] Andrea Lops, Nicola Cipriani, Gabriele Colapinto, Marcantonio de Candia, Vito Di Bari, Mauro Foglia, Gledjan Meta, Davide Savoia, Antonio Ferrara, Daniele Malitesta, Tommaso Di Noia, and Fedelucio Narducci. 2026. Distribution-Robust Reranking for Conversational Music Recommendation: PoliBaJukebox at the ACM RecSys Challenge 2026. In *Proceedings of the RecSys Challenge Workshop at the 20th ACM Conference on Recommender Systems (RecSys Challenge '26)*. ACM, Minneapolis, MN, USA. doi:10.1145/3842413.3842418

[13] Maxime Manderlier and Fabian Lecron. 2026. Picking is Not Ranking, and Explanation Quality Has Many Dimensions: Lessons for Conversational Music Recommendation. In *Proceedings of the RecSys Challenge Workshop at the 20th ACM Conference on Recommender Systems (RecSys Challenge '26)*. ACM, Minneapolis, MN, USA. doi:10.1145/3842413.3842426

[14] Sergio Oramas, Andres Ferraro, Alvaro Sarasua, and Fabien Gouyon. 2024. Talking to Your Recs: Multimodal Embeddings for Recommendation and Retrieval. In *2nd Music Recommender Systems Workshop (MuRS) at ACM RecSys*.

[15] Enrico Palumbo, Gustavo Penha, Andreas Damianou, José Luis Redondo García, Timothy Christopher Heath, Alice Wang, Hugues Bouchard, and Mounia Lalmas. 2025. Text2Tracks: Prompt-based Music Recommendation via Generative Retrieval. *CoRR* abs/2503.24193 (2025). doi:10.48550/arXiv.2503.24193

[16] Madhav Patil, Shreyas Biradar, and Divyansh Bhatia. 2026. Routing Discovery in Conversational Music Recommendation. In *Proceedings of the RecSys Challenge Workshop at the 20th ACM Conference on Recommender Systems (RecSys Challenge '26)*. ACM, Minneapolis, MN, USA. doi:10.1145/3842413.3842421

[17] Nidhin Pattaniyil, Semih Yagli, and Tanwir Zaman. 2026. State-Driven Retrieval and Learned Re-Ranking for Conversational Music Recommendation. In *Proceedings of the RecSys Challenge Workshop at the 20th ACM Conference on Recommender Systems (RecSys Challenge '26)*. ACM, Minneapolis, MN, USA. doi:10.1145/3842413.3842420

[18] Aditya Rai, Venkatesh Shukla, Sanskar Aggarwal, and Amandeep Kaur. 2026. Dialogue-Aware Conversational Music Recommendation via Multi-Modal Retrieval and Weighted RRF. In *Proceedings of the RecSys Challenge Workshop at the 20th ACM Conference on Recommender Systems (RecSys Challenge '26)*. ACM, Minneapolis, MN, USA. doi:10.1145/3842413.3842417

[19] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI*. AUAI Press.

[20] Markus Schedl, Stefan Brandl, Oleg Lesota, Emilia Parada-Cabaleiro, David Penz, and Navid Rekabsaz. 2022. LFM-2b: A Dataset of Enriched Music Listening Events for Recommender Systems Research and Fairness Analysis. In *CHIIR*. ACM, 337–341. doi:10.1145/3498366.3505791

[21] Suryaa Veerabathiran Seran. 2026. Dialogue-Aware Music Recommendation via Fused Retrieval and Learned Ranking for the TalkPlayData Challenge. In *Proceedings of the RecSys Challenge Workshop at the 20th ACM Conference on Recommender Systems (RecSys Challenge '26)*. ACM, Minneapolis, MN, USA. doi:10.1145/3842413.3842431

[22] Youness Soussou, Loubna Mekouar, and Youssef Iraqi. 2026. A Multi-Source Retrieve–Fuse–Rerank System for Conversational Music Recommendation. In *Proceedings of the RecSys Challenge Workshop at the 20th ACM Conference on Recommender Systems (RecSys Challenge '26)*. ACM, Minneapolis, MN, USA. doi:10.1145/3842413.3842425

[23] Simran Sundrani and Mohan Bhambhani. 2026. MiniMaestro: Resource-Conscious Conversational Music Recommendation with a Single Open-Weight 8B Model. In *Proceedings of the RecSys Challenge Workshop at the 20th ACM Conference on Recommender Systems (RecSys Challenge '26)*. ACM, Minneapolis, MN, USA. doi:10.1145/3842413.3842428

[24] Sanjeev Suresh. 2026. When the Label Ignores the Request: Auditing Policy-Selected Targets in Synthetic Conversational Music Recommendation. In *Proceedings of the RecSys Challenge Workshop at the 20th ACM Conference on Recommender Systems (RecSys Challenge '26)*. ACM, Minneapolis, MN, USA. doi:10.1145/3842413.3842427

[25] Tomoya Terai. 2026. Auditing Alignment among Item Relevance, Goal Progress, and LLM-Judged Responses in Music-CRS. In *Proceedings of the RecSys Challenge Workshop at the 20th ACM Conference on Recommender Systems (RecSys Challenge '26)*. ACM, Minneapolis, MN, USA. doi:10.1145/3842413.3842423

[26] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *CoRR* abs/2502.14786 (2025). doi:10.48550/arXiv.2502.14786

[27] Marco Valentini, Bianca Di Bitetto, Gianmichele De Palma, Berardino Como, Roberta Russo, Francesco Maria Dicataldo, Francesco Salvatore Desiderato, Francesco Falcone, Chiara Mallamaci, Antonio Ferrara, Daniele Malitesta, Tommaso Di Noia, and Fedelucio Narducci. 2026. Beyond Single-Signal Retrieval: A Graph-Aware Agentic Framework for Conversational Music Recommendation. In *Proceedings of the RecSys Challenge Workshop at the 20th ACM Conference on Recommender Systems (RecSys Challenge '26)*. ACM, Minneapolis, MN, USA. doi:10.1145/3842413.3842415

[28] Artem Volgin, Ben orli Nathanson, Daniele Dal Grande, Alisa Milchevskaya, Adham Al Hossary, and Rami Albatal. 2026. Frozen Embeddings, LLM Retrieval Lanes, Critic-Guided Replies: Conversational Music Recommendation at an Information-Limited Ceiling. In *Proceedings of the RecSys Challenge Workshop at the 20th ACM Conference on Recommender Systems (RecSys Challenge '26)*. ACM, Minneapolis, MN, USA. doi:10.1145/3842413.3842424

[29] Ryohei Wakatsuki. 2026. A Practical Multi-Source Pipeline for Conversational Music Recommendation in the RecSys Challenge 2026. In *Proceedings of the RecSys Challenge Workshop at the 20th ACM Conference on Recommender Systems (RecSys Challenge '26)*. ACM, Minneapolis, MN, USA. doi:10.1145/3842413.3842416

[30] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *ICASSP*. IEEE.

[31] Sungwook Yoo and Sewook Yoo. 2026. Two Views, One Voice: Evidence-Grounded Conversational Music Recommendation. In *Proceedings of the RecSys Challenge Workshop at the 20th ACM Conference on Recommender Systems (RecSys Challenge '26)*. ACM, Minneapolis, MN, USA. doi:10.1145/3842413.3842419

[32] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *CIKM*. ACM, 177–186.

[33] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *CoRR* abs/2506.05176 (2025). doi:10.48550/arXiv.2506.05176

[34] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS Datasets and Benchmarks Track)*.