TalkPlayData 2: An Agentic Synthetic Data Pipeline for Multimodal Conversational Music Recommendation

Anonymous ACL submission

Abstract

We introduce TalkPlayData 2, a multimodal conversational music recommendation dataset constructed through an agentic synthetic generation pipeline. The pipeline employs four specialized sub-agents across two stages. First, a Listener Profile LLM and a Conversation Goal LLM condition each session on a rich user profile and a conversational goal. Second, a Listener LLM and a Recsys LLM then role-play under asymmetric information – the listener knows the goal but not the recommendation pool, while the recsys knows the pool but not the goal. All four agents jointly process audio previews, album artwork, lyrics, metadata, and semantic tags, and emit per-turn reasoning and preference feedback alongside each utterance. The resulting dataset contains 16.2K conversations with 8-turn dialogues, listener profiles, structured goals, reasoning annotations, and cold-start splits. Human evaluation rates TalkPlayData 2 highest in both relevance (4.11) and naturalness (4.15) on a 5-point Likert scale, surpassing the human-curated dataset (4.08 / 4.01) and single-LLM synthetic baselines. Diversity analysis further shows a 2.8× increase in user-utterance Vendi Score over the strongest single-LLM baseline (50.16 vs. 18.13) and an order-of-magnitude reduction in topic-distribution KL divergence from uniform (0.110 vs. 1.639), demonstrating that agentic, multimodal conditioning systematically counteracts mode collapse.[1]

1 Introduction

Conversational recommendation systems provide recommendations through a natural language dialog with users, requiring both multi-turn recommendation capabilities and natural language response generation (Goker and Thompson, 2000; Christakopoulou et al., 2016; Zhang et al., 2018). At each turn, these systems predict ranked lists of relevant music tracks based on conversation history and user queries while understanding preferences, context, and diverse query types. Beyond recommendation, systems generate engaging responses that describe recommendations and assist users in music exploration through natural language explanations (Doh et al., 2024).

The main blockers for developing such a system may have been the lack of large-scale and high-quality datasets. For example, CPCD, a human-curated dataset, consists of only 917 conversations in total (Chaganty et al., 2023). Recent studies have been proposed to address this issue by actively adopting large language models (LLMs) (Leszczynski et al., 2023; Doh et al., 2024, 2025). Specifically, these approaches generate conversation trajectories by feeding track sequences into LLMs alongside fixed task prompts.

While the recent methods and their datasets have initiated developing and evaluating conversational music recommendation systems, there is room for improvements on various aspects. First, relying on single LLM with static prompts restricts diversity and runs the risk of mode collapse (Zhang et al., 2025; Schaffelder and Gatt, 2025; Wang et al., 2025; Chen et al., 2025), where the generated dialogues merely replicate the same recommendation scenario. Second, none of these frameworks incorporate multimodal capabilities, failing to reflect real-world music consumption where listeners engage with audio, metadata, lyrics, semantic tags, and album art simultaneously. Finally, these datasets lack rich conversation annotation, such as explicit user preferences (Liu et al., 2010) or the underlying reasoning steps (Wei et al., 2022) behind recommendations, both of which are indispensable components in modern recommendation systems.

To address these limitations, we propose an agentic synthetic data generation pipeline and the resulting dataset, TalkPlayData 2. The pipeline employs

[1] TalkPlayData 2and its generation code will be open-sourced after the review stage.

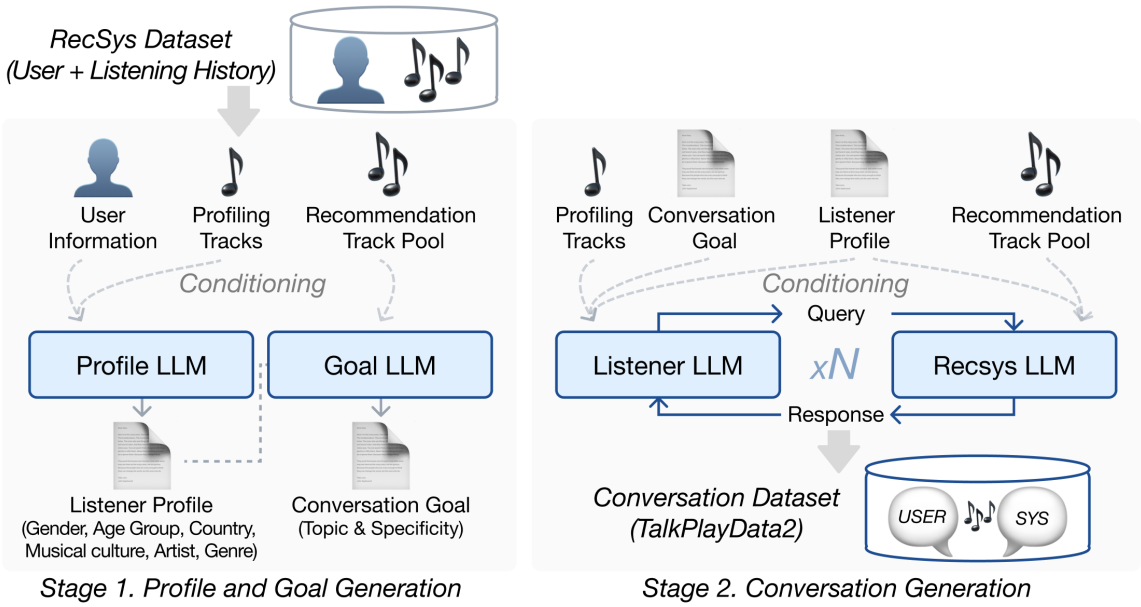

Figure 1: Overview of TalkPlayData 2 pipeline, consisting of four LLMs with specialized roles.

four specialized LLMs in two stages (illustrated in Figure 1). In the first stage, a Listener Profile LLM infers rich user profile from listening history and demographic information, while a Conversation Goal LLM synthesizes a concrete conversation goal from target recommendation track pools and listener profile. Together, these two outputs provide rich conditioning for each conversation, capturing both *who the listener is* and *what they aim to achieve*. In the second stage, a Listener LLM and a RecSys LLM conduct a Wizard-of-Oz-style (Dinan et al., 2018) role-play under asymmetric information: the Listener LLM is provided with the conversation goal but not the recommendation pool, whereas the Recsys LLM has access to the recommendation pool but not the goal. As a result, the generated conversations authentically reflect the dynamics of real-world music recommendation interactions, where users have specific goals but the system must infer them through dialogue. Moreover, beyond surface-level queries and responses, the pipeline also captures per-turn reasoning steps from both agents and preference annotations from the listener, yielding a richly annotated dataset. Our main contributions are as follows:

Multi-agent pipeline for diverse conversations. By decomposing data generation into four specialized LLM roles, each conditioned on listener profiles and structured conversation goals, the pipeline systematically prevents mode collapse and produces conversations spanning diverse topics and query specificities.

Multimodal content grounding. All four LLMs process audio previews, album artwork, textual metadata, lyrics, and semantic tags simultaneously, enabling conversations that authentically reflect how listeners engage with music across modalities.

Rich annotations for conversation recommender. Each conversation in TalkPlayData 2 is accompanied by a listener profile, a structured conversation goal, per-turn reasoning from both the Listener and Recsys LLMs, and listener preference labels indicating whether each recommendation advances the conversation goal.

2 Core Idea

In this section, we introduce a simplified formulation of the creation process of TalkPlayData 2, $y = f_\theta(x_{profile}, x_{goal}, x_{music})$. f is the agentic pipeline, θ indicates the model weights of the involved LLM, $x_{profile}$ is a listener's demographic information and preferences, x_{goal} represents the conversation objectives and scenarios, and x_{music} is a set of music data, covering text, audio, and image modalities. y indicates the outcome, a multi-turn conversation between the listener and the recommendation system about music discovery, consisting of queries, music items, and responses.

2.1 Grounding Music Data

To achieve factual data generation, TalkPlayData 2 is primarily based on a source dataset for the details about music items, x_{music}. In other words, y_{music}, the recommended tracks, must be

merely a result of sequential selections of x_{music}, the recommendation pool. More details about the source data are provided in subsection 3.1. The music data x_{music} is a set of loosely relevant music items, e.g., music tracks in a listening session in this paper (equivalent to playlists as in (Doh et al., 2025)). It is provided with a rich set of metadata, tags, lyrics, audio, and album art images.

2.2 Data Generation Agents

The creation system f is formed by multiple LLM sessions, i.e., a set of agents, each of which has a role – as a listener profiler (Profile LLM, $g_{profile}$), a conversation goal setter (Goal LLM, g_{goal}), a listener (Listener LLM, $g_{listener}$), and a recommendation system (Recsys LLM, g_{recsys}). This agentic approach presents several advantages compared to relying on a single LLM session such as (Doh et al., 2025).

Most critically, it systematically prevents the agent from 'cheating' by looking at the data provided to other roles. This makes the data generation process highly realistic. For example, a conversation goal is shared with the Listener LLM so that it generates relevant queries to the Recsys LLM and achieves its goal. However, the goal is not shared with the Recsys LLM, whose role is to guess the goal of the Listener LLM through conversation. This approach also provides the role, task, and behavior instructions to each LLM with high clarity, leading to generate high-quality data. This is well-aligned with the recent multi-agent systems such as Fellowship of the LLMs (Arif et al., 2024), MAG-V (Sengupta et al., 2024), and AgentSGEN (Xuan et al., 2025), where dedicated roles improved semantic fidelity and reduce mode collapse.

2.3 User Profile and Conversation Goal

Although the LLM generation process is often stochastic, it is well-known that naively sampling multiple times does not lead to diversifying the generation outcomes. Rather, a mode collapse often occurs, where the generated texts become too similar to each other in style and logic (Zhang et al., 2025; Schaffelder and Gatt, 2025; Wang et al., 2025; Chen et al., 2025). We observed this in our preliminary experiment - even when using different music items, the generated conversations had very similar styles. To address this issue, we utilize user demographic information and pre-defined conversation goals to generate diverse conversations. Furthermore, utilizing Goal LLM and Profile LLM, we enhance $x_{profile}$ and x_{goal} to be more suitable for recommending music x_{music}, enabling the generation of more diverse and realistic conversations. The details are provided in subsection 3.2.

3 The Creation Steps of TalkPlayData 2

3.1 Overview

Base Dataset The foundation of TalkPlayData 2 is the LFM-2b dataset ((Schedl et al., 2022)), which provides session-based music listening history data of over 120,000 users spanning more than 15 years (February 2005 to March 2020). Beyond basic metadata (track, album, artist name), LFM-2b provides rich additional information including user demographic data (country, gender, age), last.fm genre/style annotations, and ID mappings to Spotify track identifiers. Additional multimodal information is acquired through the provided Spotify track identifiers and the API – preview audio snippets, album art images, release dates, and popularity metrics (as of 2025 July). Finally, we used pretrained music information retrieval models to estimate rich information: Madmom (Böck et al., 2016) for tempo, key, and chords as well as Whisper (Radford et al., 2023) for lyrics.

Data Split To reflect real-world recommendation scenarios, we performed a chronological data split. Sessions after 2019 were reserved for testing, while earlier sessions were used for training. To create the multimodal conversation dataset, we filter listening history sessions to include only tracks with Spotify track identifier mappings. Furthermore, to address cold-start user and cold-start item scenarios, we carefully sampled the test set conversations. Out of 1,000 test set conversations, 800 were sampled from the warm user pool, while 200 were sampled from the cold user pool. During sampling, we ensured balanced sampling across demographic attributes - country, gender, and age groups. This balanced sampling helps evaluate the model's performance across diverse user segments. Detailed statistics are provided in section 4.

Large Language Models The Google Gemini 2.5 Flash (Comanici et al., 2025) is chosen to create TalkPlayData 2. Google Gemini is the only available LLM API that supports the three modalities of TalkPlayData 2, and their models have demonstrated strong music understanding capabilities in various benchmarks (Ghosh et al., 2025; Carone et al., 2025a,b; Lee et al., 2025; He et al., 2025; Kumar et al., 2025; Ma et al., 2025).

Algorithm 1 Data Generation Process for Multi-modal Music Recommendation Conversations

Require: Listening sessions S, Set of tracks M, User Profile U, Conversation Goal G, Query message Q, Recommended Music M_t, Response R, Thought T, Progress Towards Goal P

1: **for** each listening session s with $|M| \geq 21$ tracks **do**
2: $M_{profile} \leftarrow$ sample 5 tracks of M
3: $M_{pool} \leftarrow$ sample 16-32 tracks of $M - M_{profile}$
4: $U_{base} \leftarrow$ user demographic information of s (age, country, gender)
5: $G_{base} \leftarrow$ sample 3 templates from goal dictionary
6: $U_{final} \leftarrow$ ListenerProfileLLM($M_{profile}$, U_{base})
7: $G_{final} \leftarrow$ ConversationGoalLLM(G_{base}, M_{pool}, U_{final})
8: $Q_1 \leftarrow$ ListenerLLM(U_{final}, G_{final}, $M_{profile}$) ▷ First query message (Q_1)
9: $T_1^t, M_1, R_1 \leftarrow$ RecsysLLM(U_{final}, M_{pool}, Q_1) ▷ First track (M_1), response (R_1)
10: $conversation \leftarrow [Q_1, M_1, R_1]$
11: **for** turn $t = 2$ to 8 **do**
12: $P_t, T_t^l, Q_t \leftarrow$ ListenerLLM(U_{final}, G_{final}, $M_{profile}$, $conversation$)
13: $conversation$.append(Q_t)
14: $T_t^r, M_t, R_t \leftarrow$ RecsysLLM(U_{final}, M_{pool}, $conversation$)
15: $conversation$.append(M_t R_t)
16: **end for**
17: **end for**

3.2 Generation Process

The overall process iterates over the listening sessions S and converts each session $s \in S$ into a conversation. Each session s consists of a list of tracks M and basic user demographic information U, including age group, gender, and country. For each conversation, we require sessions containing at least 21 tracks to ensure a sufficient recommendation pool. From each session, 5 tracks are sampled as profiling tracks ($M_{profile}$) to inform the conversation style sampling, while another 16-32 tracks form the recommendation pool (M_{pool}).

For a single conversation, four separate LLM sessions are created, each of which corresponds to $g_{profile}$, g_{goal}, $g_{listener}$, and g_{recsys}, respectively. Their instructions are carefully designed after many iterations of reviews and updates to ensure correct behaviors and response formats. As outlined in Algorithm 1 and Figure 1, the overall generation process consists of two stages: i) profiling and goal generation, and ii) conversation generation. During the first stage (lines 1-8), we create a user profile U_{final} based on profiling tracks $M_{profile}$ and demographic information U_{base}; and a conversation goal G_{final} from base goal templates G_{base}, recommendation pool M_{pool}, and user profile U_{final}. This customization is responsible for diversifying the conversation in style. The conversation generation stage (lines 11-15) follows with alternating API calls between the listener and the recommendation systems, where queries Q_t, music recommendations M_t, and responses R_t build upon the conversation history while maintaining coherence with the conversation goals G_{final}. The detailed instructions and responses for the LLMs are provided in Appendix A.

Listener Profile LLM The role of the Listener Profile LLM is to analyze the profiling tracks and infer high-level preference information of the listener, given demographic information, such as gender, age group, and country. During generating TalkPlayData 2 as well as using it, this information provides the personalization aspect, which is crucial in modern recommendation systems and music consumption (Schedl et al., 2021; Kaminskas and Ricci, 2012; North and Hargreaves, 2008). The LLM combines the provided demographic profile with the track analysis to estimate musical preferences including preferred musical culture, top artist, and top genre. All the text data (metadata, tags, and lyrics), audio, and image data described in subsection 3.1 are provided to the LLM.

Conversation Goal LLM The Conversation Goal defines the session-level goal that the listener wants to achieve through conversation with the recommendation system. To guarantee the overall diversity of the conversations in TalkPlayData 2, a diverse set of conversation goals is needed; while

Table 1: Conversation Goal Axis 1 - Topics

Code	Description	Example
A	Audio-Based Discovery	"Discover songs with immersive soundscapes"
B	Lyrical Discovery	"Songs about love"
C	Visual-Musical Connections	"Music that looks colorful and vibrant"
D	Contextual & Situational	"Music for working and studying"
E	Interactive Refinement	"Let's play hard rock and transit to modern rock"
F	Metadata-Rich Exploration	"Find multiple songs from the Hamilton musical"
G	Mood & Emotion-Based	"I need something to cheer me up"
H	Artist & Discography Discovery	"Tell me about this artist's other works"
I	Cultural & Geographic	"Music from Alaska"
J	Social & Popularity Context	"What's trending right now?"
K	Temporal & Era Discovery	"Music from the 80s please"

Table 2: Conversation Goal Axis 2 - Specificities

Code	Description	Example
LL	Low query specificity Low target specificity	"Play some chill music" (Many tracks are possible as a successful recommendation)
HL	High query specificity low target specificity	"Find bebop jazz with saxophone, 1950s-60s" (Many tracks are possible, the query is somewhat specific)
LH	Low query specificity high target specificity	"What was the popular song from a recent musical movie?" (One or few tracks are possible, the query is not specific)
HH	High query specificity high target specificity	"Windup by Hayoung Lyou, the jazz composer and pianist" (One track is possible, the query is highly specific)

each conversation goal should be plausible given the recommendation pool (Li et al., 2024). Before the generation process, a set of 44 conversation goal templates is prepared. A template is defined by two properties, the topic and the specificities. In total, 11 topics are defined to cover various types of multi-modal music discovery conversations, as listed in Table 2. These topics decide which aspect of music the conversation will be based on and cover recommendation scenarios based on audio (Deldjoo et al., 2024; Van den Oord et al., 2013), lyrics (Patra et al., 2017; Vystrčilová and Peška, 2020), visual information (Saito and Itoh, 2011; Libeks and Turnbull, 2011), and emotion ((Han et al., 2010)). The specificities define how specific the query and the target music are, resulting in 4 cases as in Table 2. This two-dimensional formalization of LL, HL, LH, and HH provides a structured view of recommendation scenarios such as exploratory search (Marchionini, 2006; Schedl et al., 2015), lookup tasks (Marchionini, 2006), and query granularity (Sun and Zhang, 2018; Jannach et al., 2021).

There are two steps to generate a conversation goal. First, three templates are randomly sampled. They decide the potential directions, but they are still template candidates, since some of them may not be plausible per the recommendation pool. For example, the recommendation pool may consist of all instrumental music, which would limit lyric-based conversations. Second, the three base conversation goals are fed to the Goal LLM (g_{goal}), whose role is to select the most plausible goal based on the recommendation pool and customize the overall goal with concrete examples.

To improve conversation pacing and realism, each conversation goal includes a target turn count that guides the expected resolution time: HH specificity goals target 1-2 turns (quick resolution), HL specificity targets 3-4 turns (moderate exploration), LL specificity targets 3-7 turns (extensive exploration), and LH specificity targets 6-8 turns (detailed exploration). The target turn count is determined by the Goal LLM (g_{goal}) based on goal complexity and recommendation pool content. While conversations always continue to the full 8 turns for consistent training data, the target turn count

Table 3: Comparison among conversational music recommendation datasets. CS refers to cold-start split. Gray text indicates single-turn datasets.

Dataset	Profile	Goal	Thought	CS	Conv.	Track	User	Turns
JAM (Melchiorre et al., 2025)	✓	✗	✗	N/A	112K	100k	104K	1.00
Text2Track (Palumbo et al., 2024)	✗	✗	✗	N/A	1M	500K	-	1.00
CPCD (Chaganty et al., 2023)	✗	✗	✗	✗	0.1K	107K	-	5.70
LP-MusicDialog (Doh et al., 2024)	✗	✗	✗	✗	288K	391K	-	4.97
TalkPlayData 1 (Doh et al., 2025)	✗	✗	✗	✓	532K	406K	-	6.95
TalkPlayData 2 (**Ours**)	✓	✓	✓	✓	16.2K	47K	9k	8.00

influences the listener's pacing strategy and goal achievement approach.

Listener LLM and Recsys LLMs Based on the profiling tracks, the conversation goal, and the listener profile, the conversation is initiated by the Listener LLM ($g_{listener}$). On Recsys LLM (g_{recsys}), after being initialized with the listener profile and the recommendation pool (but not the conversation goal), it starts to respond to the Listener LLM's initial query. In the subsequent turns (from Turn 2), the Listener LLM actively engages with the recommended music by listening to the audio samples and seeing the album artwork before formulating responses. This multimodal interaction allows the Listener LLM to provide more nuanced and informed feedback about the recommendations, to which Recsys LLM then makes subsequent recommendations. The Listener LLM also labels whether the Recsys LLM's recommendation is making positive progress towards achieving the goal. This information is expected to be used as a 'preference' signal during training recommendation models using reinforcement learning, or as an auxiliary classification target. Finally, in every turn, both the Listener and the Recsys LLMs generate 'thought' before generating their response message to each other. A 'thought' is used to analyze the input message (from the other LLM), increasing interpretability during both dataset creation and utilization.

4 Dataset Statistics and Evaluation

4.1 Statistics

Table 4 summarizes the key statistics of TalkPlayData 2, which consists of a 15,199 training set and 1,000 test set divided using chronological splitting to reflect real-world deployment scenarios. The test set includes 129 cold users and 2,982 cold tracks for evaluating cold-start scenarios, while a distinctive characteristic is the comprehensive inclusion of user queries, assistant responses, and detailed thought processes that provide valuable insights into the reasoning behind preferences and recommendations, enabling interpretable music recommendation systems.

Table 4: TalkPlayData 2 statistics

Entity	Training	Evaluation
Conversations	15199	1000
Total Users	8591	500
Warm Users	-	371
Cold Users	-	129
Total Tracks	43597	6761
Warm Tracks	-	3779
Cold Tracks	-	2982
Avg. Query Length	44.7	39.8
Avg. Response Length	36.6	44.6
Avg. Thought Length	89.3	101.7

Table 3 presents a comparison among conversational music recommendation datasets: including two closed-source single-turn conversation datasets (Palumbo et al., 2024; Melchiorre et al., 2025), a human conversation dataset (Chaganty et al., 2023), and two LLM-based synthetic datasets (Doh et al., 2024, 2025). While TalkPlayData 2 does not have a large number of conversations, it represents the dataset most similar to real music recommendation scenarios, featuring 1) long conversation turns, 2) user profiles, 3) conversation goals, 4) thought, and 5) cold-start splits.

4.2 Human Evaluation

We assess the quality of our generated data through human evaluation, focusing on two key aspects: 1) *relevance* – determining the alignment between the retrieved music items and the user query, and 2) *naturalness* – assessing the likelihood of such a conversation occurring in real life. We adhere to a mean opinion score that uses a 5-point Lik-

Table 5: Human evaluation results on conversational music datasets. Type stands for the subject of conversation. Relevance and Naturalness show Mean Opinion Scores on a 5-point Likert scale.

Datasets	Type	Agentic	Multimodal	Relevance	Naturalness
CPCD (Chaganty et al., 2023)	Human	-	-	4.08 ±0.11	4.01 ±0.11
LP-MusicDialog (Doh et al., 2024)	Synthetic	✗	✗	3.90 ±0.10	3.95 ±0.11
TalkPlayData 1 (Doh et al., 2025)	Synthetic	✗	✗	4.04 ±0.10	4.01 ±0.09
TalkPlayData 2 **(Ours)**	Synthetic	✓	✓	4.11 ±0.09	4.15 ±0.08

ert scale. A total of 26 music experts evaluated 10 randomly sampled dialogues each, resulting in 260 total ratings. For comparison models, we select open-source conversational music recommendation datasets. CPCD (Chaganty et al., 2023) is a human conversation dataset about music, and LP-MusicDialog (Doh et al., 2024) and TalkPlayData 1 (Doh et al., 2025) are synthetic conversation datasets generated by a single LLM.

As shown in Table 5, TalkPlayData 2 achieves the highest scores in both dimensions. The high relevance score demonstrates the effectiveness of our multimodal approach, where LLMs consider both audio and visual aspects of music during recommendation, leading to more accurate and contextually appropriate suggestions compared to text-only approaches (Doh et al., 2024, 2025). The strong naturalness score highlights the effectiveness of our agentic multi-LLM framework. By orchestrating interaction between the Conversation Goal LLM and the Profile LLM, the system enables naturalistic exchanges between the Listener LLM and the RecSys LLM, thereby improving both conversational coherence and user simulation. These results suggest that our approach of using multiple LLMs creates more engaging and effective conversational recommendations compared to both human conversations (Chaganty et al., 2023) and single-LLM approaches (Doh et al., 2024, 2025).

4.3 Diversity Evaluation

To ensure the generated dialogue dataset encompasses a broad spectrum of user scenario, we conduct a diversity evaluation. We evaluate the dataset across two dimensions: sentence-level semantic diversity, and conversation-level goal diversity. [2] For a comparative analysis, we benchmark our dataset against the open-source multi-turn datasets utilized in the human evaluation in subsection 4.2.

[2]We excluded lexical diversity from this study because responses in recommender systems inherently contain item metadata, making lexical diversity highly correlated with track diversity.

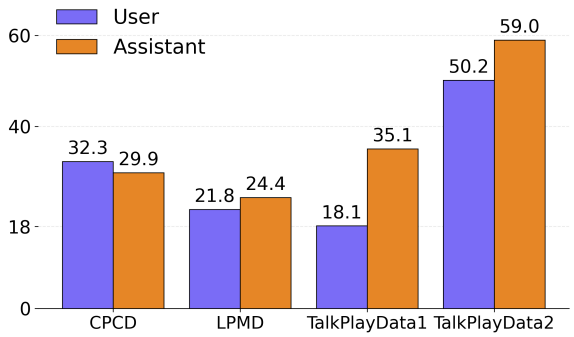

Figure 2: Vendi score result.(higher is better)

Semantic diversity We utilize the Vendi Score (Friedman and Dieng, 2022), a diversity metric based on the eigenvalues of a similarity kernel. By embedding the utterances using a pre-trained sentence encoder (`embedding-gemma-300m` (Vera et al., 2025)), the Vendi Score calculates the effective number of modes within the semantic space, allowing us to identify and penalize potential mode collapse in the synthetic data generation process. Given a normalized similarity kernel matrix $\mathbf{K}/n$ derived from n embeddings, the Vendi Score is calculated as the exponential of the Shannon entropy of its eigenvalues:

$$\text{Vendi Score}(\mathbf{K}) = \exp\left(-\sum_{i=1}^{n} \lambda_i \log \lambda_i\right)$$

where λ_i denotes the i-th eigenvalue of the normalized kernel matrix. To account for discrepancies in dataset scale, we strictly limit our evaluation to their test splits and randomly sample 2,448 utterances from each dataset, matching the size of the smallest dataset (CPCD), ensuring a size-controlled and fair comparison.

As shown in Figure 2, TalkPlayData 2 achieves the highest semantic diversity on both user (50.16) and assistant (58.98) utterances, outperforming the human-curated CPCD (32.30 / 29.86) and the single-LLM synthetic baselines LP-MusicDialog (21.75 / 24.37) and TalkPlayData-1 (18.13 / 35.08). From an *agentic* perspective, TalkPlayData-1 shows a striking user-side collapse (18.13) while

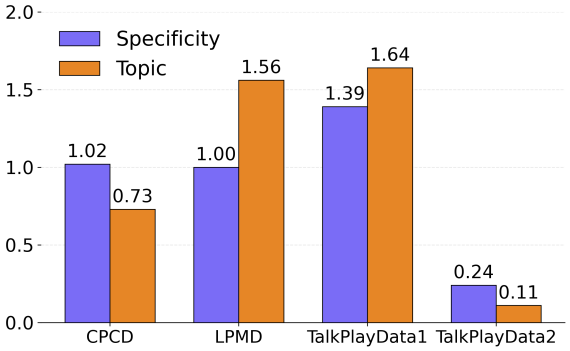

Figure 3: Kullback-Leibler divergence result.(lower is better)

retaining assistant-side variation (35.08), indicating that a single LLM with a fixed prompt converges to a narrow listener mode. Our multi-agent design conditions each conversation on a distinct profile and goal and assigns the listener and recsys roles to separate sessions with asymmetric information, lifting the user score by 2.8× and surpassing human dialogues. From a *multimodal* perspective, both text-only baselines remain below 25 on the user side because textual metadata constrains listeners to recurring references such as genre or artist, whereas grounding every agent in audio, artwork, lyrics, and tags broadens the expressible reference space and drives the largest absolute gain on the user side (+32 over TalkPlayData-1).

Goal diversity We evaluate how well the generated dataset covers various domains by comparing its empirical topic distribution against an ideal uniform distribution. We employ an LLM as a classifier, prompting Gemini-2.5-Pro (Comanici et al., 2025) to classify each conversation into one of 4 specificities and 11 predefined topics in subsection 3.2. We then use the Kullback-Leibler divergence (KLD) to measure how close the resulting empirical distributions are to a uniform distribution. A lower KLD indicates that the generated conversations are well-balanced across all categories, avoiding topical mode collapse. We note that the baseline datasets were not constructed with this taxonomy; their KLD values therefore serve as a reference point illustrating the structural effect of explicit goal steering, rather than a direct head-to-head comparison.

As shown in Figure 3, TalkPlayData 2 attains lower KLD values on both specificity (0.240) and topic (0.110) compared to TalkPlayData-1 (1.39 / 1.64), LP-MusicDialog (1.00 / 1.56), and CPCD (1.02 / 0.73). Since the baselines were not designed with conversation goal, the comparison should be interpreted with caution. Nevertheless, a notable observation is that single-LLM unimodal approaches (TalkPlayData-1 and LP-MusicDialog) show substantially higher KLD on the topic axis (1.64 and 1.56, respectively) compared to the human-curated CPCD (0.73), suggesting that relying on a single LLM without explicit goal conditioning makes it difficult to even match the topical diversity naturally present in human conversations. Our multi-agent, goal-conditioned approach substantially reduces this gap.

5 Conclusion

In this paper, we introduced TalkPlayData 2, a new multimodal dataset for conversational recommendation systems. In the data generation pipeline, separate LLM calls are first made to create a listener profile and a conversation goal for each conversation. Using them as a condition, two separate LLMs talk to each other under the role of a music listener and a music recommendation system. The conversation is conducted for 8 turns, and the data is collected as a conversation. Notably, all the LLMs are multimodal, enabling to generate conversation with multimodal aspects of music being considered. The LLMs have access to different subsets of the information, a design choice that is highly similar to the real-world conversational recommendation systems. In the evaluation, we conducted a diversity evaluation as well as a human evaluation, which shows that TalkPlayData 2is a promising dataset for training and evaluating conversational music recommendation systems.

There are still many interesting directions to explore in the future. First, the in-context recommendation of the Recsys LLM has a limitation in the number of tracks it can consider. Expanding its recommendation pool size is a natural direction. Second, although TalkPlayData 2 consists of highly natural conversations, it is still limited in various aspects including the speaking style and language. Third, due to the cost of the LLMs, during the data generation, only a short audio snippet and a small album cover image are used. Using longer audio and more diverse visual information (such as music videos, live performances, and any other modalities) can make the data even more deeply multimodal, enabling holistic multimodal conversational recommendation systems.

Limitations

While TalkPlayData 2 advances multimodal conversational music recommendation, several limitations remain.

Limited Domain Scope The proposed pipeline is demonstrated solely for music recommendation. Although the multi-agent design and multimodal grounding are domain-agnostic in principle, their effectiveness in other conversational recommendation domains (e.g., movies, books, or products) remains unvalidated. Generalizing the framework to other domains would require redesigning the goal taxonomy and sourcing modality-appropriate data.

Fixed Vocabulary of Conversation Goals The closed set of 44 goal templates (11 topics $\times$ 4 specificity levels) establishes structured diversity but cannot cover the full range of real-world user intents. In practice, user goals are often multi-faceted and emergent, crystallizing gradually through dialogue rather than being predetermined. This fixed taxonomy inherently under-represents long-tail intent patterns and complex conversational shifts that fall outside the predefined categories.

Diversity Evaluation Coverage The Vendi Score measures semantic diversity at the utterance level but does not capture structural diversity of conversational trajectories—for instance, recurring turn-taking patterns, goal-resolution pacing, or the variety of multi-turn reasoning chains. A more comprehensive diversity evaluation would require trajectory-level metrics that assess the structural variability of full conversations, not just individual utterances.

References

Samee Arif, Sualeha Farid, Abdul Hameed Azeemi, Awais Athar, and Agha Ali Raza. 2024. The fellowship of the llms: Multi-agent workflows for synthetic preference optimization dataset generation. *Preprint*, arXiv:arXiv:2408.08688.

Sebastian Böck, Filip Korzeniowski, Jan Schlüter, Florian Krebs, and Gerhard Widmer. 2016. Madmom: A new python audio and music signal processing library. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1174–1178.

Brandon James Carone, Iran R Roman, and Pablo Ripollés. 2025a. Evaluating multimodal large language models on core music perception tasks. *arXiv preprint arXiv:2510.22455*.

Brandon James Carone, Iran R Roman, and Pablo Ripollés. 2025b. The muse benchmark: Probing music perception and auditory relational reasoning in audio llms. *arXiv preprint arXiv:2510.19055*.

Arun Tejasvi Chaganty, Megan Leszczynski, Shu Zhang, Ravi Ganti, Krisztian Balog, and Filip Radlinski. 2023. Beyond single items: Exploring user preferences in item sets with the conversational playlist curation dataset. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2754–2764.

Jianhao Chen, Zishuo Xun, Bocheng Zhou, Han Qi, Hangfan Zhang, Qiaosheng Zhang, Yang Chen, Wei Hu, Yuzhong Qu, Wanli Ouyang, and Shuyue Hu. 2025. Do we truly need so many samples? multi-llm repeated sampling efficiently scales test-time compute. *Preprint*, arXiv:2504.00762.

Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 815–824.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Yashar Deldjoo, Markus Schedl, and Peter Knees. 2024. Content-driven music recommendation: Evolution, state of the art, and challenges. *Computer Science Review*, 51:100618.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

SeungHeon Doh, Keunwoo Choi, Daeyong Kwon, Taesu Kim, and Juhan Nam. 2024. Music discovery dialogue generation using human intent analysis and large language models. *arXiv preprint arXiv:2411.07439*.

Seungheon Doh, Keunwoo Choi, and Juhan Nam. 2025. Talkplay: Multimodal music recommendation with large language models. *arXiv preprint arXiv:2502.13713*.

Dan Friedman and Adji Bousso Dieng. 2022. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*.

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sanggil Lee, Zhifeng Kong, Joao Felipe Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, and 1 others. 2025. Music flamingo: Scaling music understanding in audio language models. *arXiv preprint arXiv:2511.10289*.

M Goker and Cynthia Thompson. 2000. The adaptive place advisor: A conversational recommendation system. In *Proceedings of the 8th German workshop on case based reasoning*, pages 187–198.

Byeong-jun Han, Seungmin Rho, Sanghoon Jun, and Eenjun Hwang. 2010. Music emotion classification and context-based music recommendation. *Multimedia Tools and Applications*, 47(3):433–460.

Peize He, Zichen Wen, Yubo Wang, Yuxuan Wang, Xiaoqian Liu, Jiajie Huang, Zehui Lei, Zhuangcheng Gu, Xiangqi Jin, Jiabing Yang, and 1 others. 2025. Audiomarathon: A comprehensive benchmark for long-context audio understanding and efficiency in audio llms. *arXiv preprint arXiv:2510.07293*.

Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)*, 54(5):1–36.

Marius Kaminskas and Francesco Ricci. 2012. Contextual music information retrieval and recommendation: State of the art and challenges. *Computer Science Review*, 6(2-3):89–119.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, and 1 others. 2025. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *arXiv preprint arXiv:2508.13992*.

Kuan-Yi Lee, Tsung-En Lin, and Hung-Yi Lee. 2025. Audio-maestro: Enhancing large audio-language models with tool-augmented reasoning. *arXiv preprint arXiv:2510.11454*.

Megan Leszczynski, Shu Zhang, Ravi Ganti, Krisztian Balog, Filip Radlinski, Fernando Pereira, and Arun Tejasvi Chaganty. 2023. Talk the walk: synthetic data generation for conversational music recommendation. *arXiv preprint arXiv:2301.11489*.

Chuang Li, Yang Deng, Hengchang Hu, Min-Yen Kan, and Haizhou Li. 2024. Incorporating external knowledge and goal guidance for llm-based conversational recommender systems. *Preprint*, arXiv:2405.01868.

Janis Libeks and Douglas Turnbull. 2011. You can judge an artist by an album cover: Using images for music annotation. *IEEE MultiMedia*, 18(4):30–37.

Nathan N Liu, Evan W Xiang, Min Zhao, and Qiang Yang. 2010. Unifying explicit and implicit feedback for collaborative filtering. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1445–1448.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, and 1 others. 2025. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*.

Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46.

Alessandro B Melchiorre, Elena V Epure, Shahed Masoudian, Gustavo Escobedo, Anna Hausberger, Manuel Moussallam, and Markus Schedl. 2025. Just ask for music (jam): Multimodal and personalized natural language music recommendation. *arXiv preprint arXiv:2507.15826*.

Adrian North and David Hargreaves. 2008. *The social and applied psychology of music*. OUP Oxford.

Enrico Palumbo, Gustavo Penha, Andreas Damianou, José Luis Redondo García, Timothy Christopher Heath, Alice Wang, Hugues Bouchard, and Mounia Lalmas. 2024. Text2tracks: Generative track retrieval for prompt-based music recommendation. In *The 1st Workshop on Risks, Opportunities, and Evaluation of Generative Models in Recommender Systems (ROEGEN@ RECSYS'24)*.

Braja Gopal Patra, Dipankar Das, and Sivaji Bandyopadhyay. 2017. Retrieving similar lyrics for music recommendation system. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 290–297.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Yuri Saito and Takayuki Itoh. 2011. Musicube: a visual music recommendation system featuring interactive evolutionary computing. In *Proceedings of the 2011 Visual Information Communication-International Symposium*, pages 1–6.

Max Schaffelder and Albert Gatt. 2025. Synthetic eggs in many baskets: The impact of synthetic data diversity on llm fine-tuning. *arXiv preprint arXiv:2511.01490*.

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 *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, pages 337–341.

Markus Schedl, Peter Knees, Brian McFee, and Dmitry Bogdanov. 2021. Music recommendation systems: Techniques, use cases, and challenges. In *Recommender systems handbook*, pages 927–971. Springer.

Markus Schedl, Peter Knees, Brian McFee, Dmitry Bogdanov, and Marius Kaminskas. 2015. Music recommender systems. *Recommender systems handbook*, pages 453–492.

Saptarshi Sengupta, Harsh Vashistha, Kristal Curtis, Akshay Mallipeddi, Abhinav Mathur, Joseph Ross, and Liang Gou. 2024. Mag-v: A multi-agent framework for synthetic data generation and verification. *Preprint*, arXiv:arXiv:2412.04494.

Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 235–244.

Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen. 2013. Deep content-based music recommendation. *Advances in neural information processing systems*, 26.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, and 1 others. 2025. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*.

Michaela Vystrčilová and Ladislav Peška. 2020. Lyrics or audio for music recommendation? In *Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics*, pages 190–194.

Tianchun Wang, Zichuan Liu, Yuanzhou Chen, Jonathan Light, Haifeng Chen, Xiang Zhang, and Wei Cheng. 2025. Diversified sampling improves scaling llm inference. *Preprint*, arXiv:2502.11027.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Vu Dinh Xuan, Hao Vo, David Murphy, and Hoang D. Nguyen. 2025. Agentsgen: Multi-agent llm in the loop for semantic collaboration and generation of synthetic data. *Preprint*, arXiv:2505.13466.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R Tomz, Christopher D Manning, and Weiyan Shi. 2025. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity. *arXiv preprint arXiv:2510.01171*.

Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th acm international conference on information and knowledge management*, pages 177–186.

A API Specifications of the LLMs During Generation

During the generation process, the LLM calls consist of many long prompts, defining the task, behavior, input data, and the response format. In this Appendix, we provide a summary of the prompt as follows.

A.1 Listener Profile LLM

Input: demographics + 5 profile tracks

```
[f"You are an expert in music and
demographic analysis. Given the demographic
profile below and tracks, please analyze the
tracks and infer the most representative
preferred_musical_culture, artist and genre
that define this listener's taste.",
Demographic Profile:
- age_group: [factual age group]
- country: [factual country]
- gender: [factual gender]
- preferred_language: [factual language]
"Title: [track 1 title], Artist: [track 1
artist], ...", AudioContent, ImageContent,
. . . ,
"Title: [track 5 title], Artist: [track 5
artist], ...", AudioContent, ImageContent]
```

Output: Listener Profile (YAML)

```
preferred_musical_culture: [representative
culture]
top_1_artist: [representative artist]
top_1_genre: [representative genre]
```

A.2 Conversation Goal LLM

Input: 32 pool tracks + goal templates

```
[f"You are an expert in music listening ...
Step 1: Analyze the tracks and the provided
conversation goal templates, and select the
most appropriate one ... Step 2: Generate a
new goal that is more specific to the tracks,
based on the selected goal.",
"Title: [track 1 title], Artist: [track 1
artist], ...", AudioContent, ImageContent,
. . . ,
"Title: [track 32 title], Artist: [track 32
artist], ...", AudioContent, ImageContent,
f"Here are the conversation goal templates,
based on which you will generate the new
conversation goal:
{three_conversation_goal_templates}"]
```

Output: Conversation Goal (YAML, abridged)

```
category_code: [topic code among A-K]
specificity_code: [one of LL, HL, LH, HH]
target_turn_count: [1-8 based on
specificity]
listener_goal: [customized goal description]
listener_expertise: [description of
expertise]
initial_query_example_1: [plausible initial
query]
```

A.3 Listener LLM (First Turn)

Input: profile + goal + 5 profile tracks

```
[f"You are an AI assistant role-playing as a
music listener. Your personality, knowledge,
and objectives are STRICTLY defined by the
Listener Profile and Conversation Goal
provided below. ... For your very first
message (Turn 1), choose one of the initial
query examples provided in the Conversation
Goal and use it ...",
"Title: [profile track 1 title], ...",
AudioContent, ImageContent,
. . . ,
"Title: [profile track 5 title], ...",
AudioContent, ImageContent,
f"{listener_profile}",
f"{conversation_goal}",
f"You are starting a new music discovery
conversation. ... Turn 1. Now, create the
first turn query to RecSys ..."]
```

Output: First Listener Message (YAML)

```
thought: [internal reasoning about goal and
approach]
message: [natural opening message to the
recsys]
```

A.4 Recsys LLM (First Turn)

Input: profile + 32 pool tracks + listener msg

```
[f"... You are TalkPlay, an expert music
recommendation system with deep musical
knowledge, audio analysis capabilities, and
image analysis capabilities. ... You MUST
recommend ONLY from the provided available
tracks. ... Make personalized music
recommendations ... {listener_profile}",
"Title: [pool track 1 title], ID: [pool track
1 id], ...", AudioContent, ImageContent,
. . . ,
"Title: [pool track 32 title], ID: [pool
track 32 id], ...", AudioContent,
ImageContent,
"... Turn 1. ... Listener's message:
{listener_message} ..."]
```

Output: Recsys Response (YAML)

```
thought: [analysis of request and selection
reasoning]
track_id: [selected track id from the pool]
message: [natural response with track info
and explanation]
```

A.5 Listener LLM (Subsequent Turns)

Input: recommended track + recsys msg

```
[f"Title: [recommended title], Artist:
[recommended artist], ...", AudioContent,
ImageContent,
f"You just listened to this recommended
track: ... The recommendation system said:
'{recsys_message}' ... Assess whether this
track moves you toward achieving your
Conversation Goal ..."]
```

Output: Listener Response (YAML)

```
thought: [internal evaluation of track and
strategy]
goal_progress_assessment: [MOVES_TOWARD_GOAL
or DOES_NOT_MOVE_TOWARD_GOAL]
message: [feedback and next request toward
the goal]
```

A.6 Recsys LLM (Subsequent Turns)

Input: used track ids + listener msg

```
[f"... Previous Tracks: {used_track_ids} ...
Listener's message: '{listener_message}'
...",
"... NO DUPLICATES ... Maintain conversation
coherence and respond naturally."]
```

Output: Recsys Response (YAML)

```
thought: [analysis of feedback and next
strategy]
track_id: [next selected track id]
message: [response with new track and
reasoning]
```

B User Demographic Analysis

The dataset exhibits diverse demographic shown in Table 5. The test set includes users from 53 different countries, and when visualizing the distribution of users across continents, we observe that Europe and the Americas are dominant despite balanced sampling efforts. In terms of age distribution, teenagers and the 20s constitute the majority of users. Gender distribution shows a nearly uniform pattern across categories.

C Human Evaluation Survey

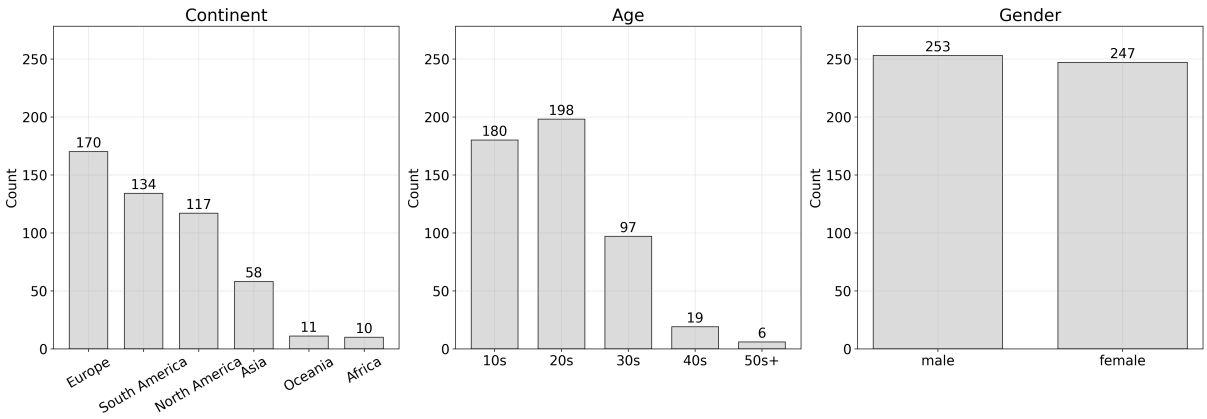

Figure 4: User Distribution of TalkPlay2 Test Dataset

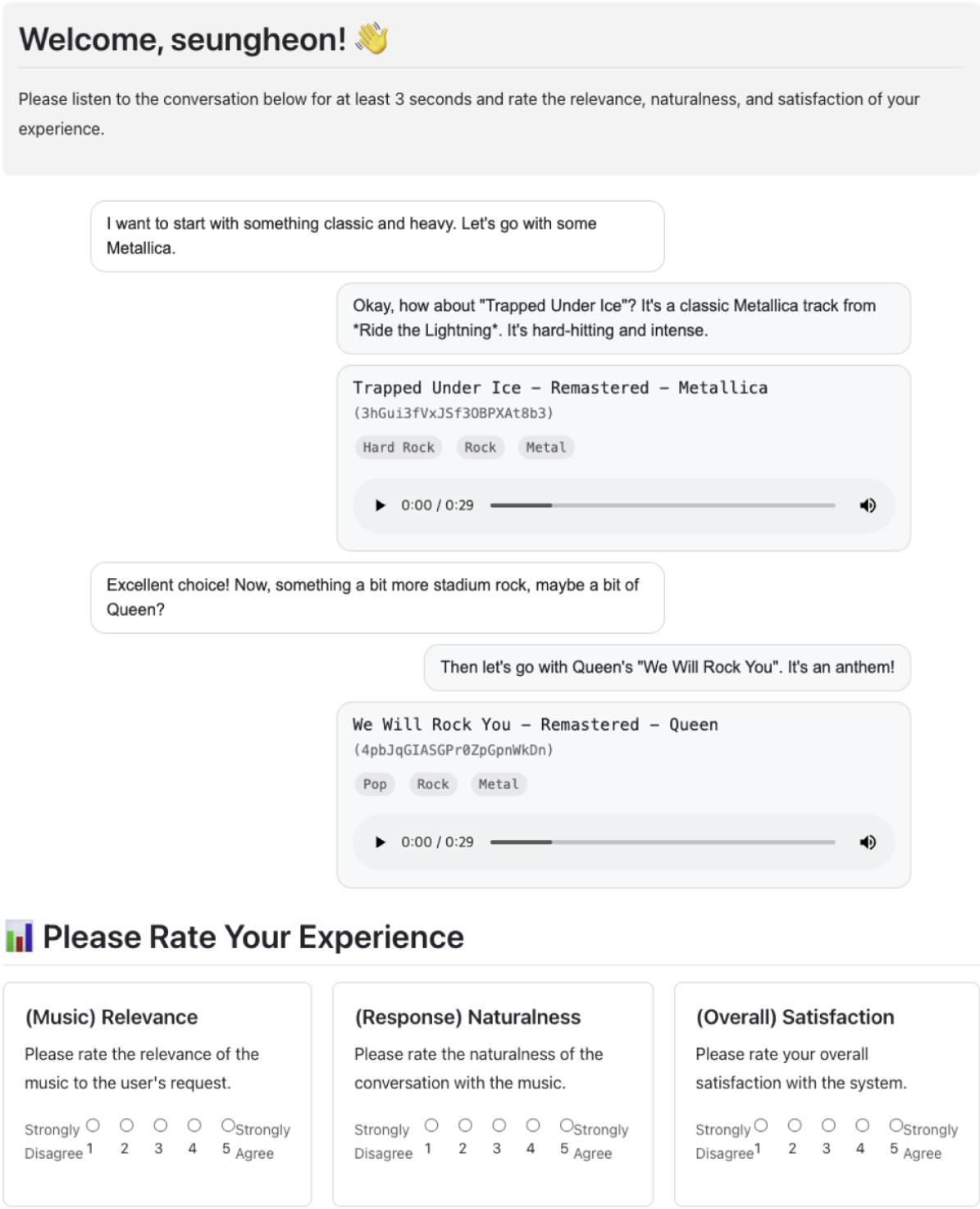

Figure 5: Survey Screenshot of Human Evaluation