arXiv:2511.16478v3 [cs.IR] 25 Aug 2026

Music Recommendation with Large Language Models: Challenges, Opportunities, and Evaluation

ELENA V. EPURE*, Deezer Research, France and Idiap Research Institute, Switzerland
YASHAR DELDJOO*, Politecnico di Bari, Italy
BRUNO SGUERRA*, Deezer Research, France
MARKUS SCHEDL, Johannes Kepler University Linz and Linz Institute of Technology, Austria
MANUEL MOUSSALLAM, Deezer Research, France

Music Recommender Systems (MRSs) have long relied on an information retrieval framing, where progress is measured mainly through accuracy on retrieval-oriented subtasks. While effective, this reductionist paradigm struggles to address the deeper question of what makes a good recommendation. Attempts to broaden evaluation, through user studies or fairness analyses, have had limited impact. The emergence of Large Language Models (LLMs) disrupts this framework: LLMs are generative rather than ranking-based, making standard accuracy metrics questionable. They also introduce challenges such as hallucinations, knowledge cutoffs, non-determinism, and opaque training data, rendering traditional train/test protocols difficult to interpret. At the same time, LLMs create new opportunities, enabling natural language (NL) interaction and even allowing models to act as evaluators.

This work argues that the shift toward LLM-driven MRSs requires rethinking evaluation. We first review how LLMs have impacted user modeling, item modeling, and NL-based recommendation in music. We then analyze evaluation practices from NLP, highlighting methodologies and open challenges relevant to MRSs. Finally, we synthesize insights, focusing on how LLM prompting applies to MRSs, to outline a structured set of success and risk dimensions. Our goal is to provide the MRSs community with an updated, pedagogical, and cross-disciplinary perspective on evaluation.

Additional Key Words and Phrases: Music Recommender Systems, Item Modeling, User Modeling, Natural Language Recommendation Requests, Large Language Models, Evaluation, Metrics, Risks, Challenges

ACM Reference Format:
Elena V. Epure, Yashar Deldjoo, Bruno Sguerra, Markus Schedl, and Manuel Moussallam. 2025. Music Recommendation with Large Language Models: Challenges, Opportunities, and Evaluation. 1, 1 (August 2025), 54 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

At its origins, the field of Music Recommender Systems (MRSs) was built upon the same foundations as the broader Recommender Systems (RSs) domain, namely, by framing its objectives as a set of information retrieval tasks. Practitioners adopting this view benefited from the rapid advancements in the RSs community and in machine learning

*Equal contribution.

Authors' Contact Information: Elena V. Epure, research@deezer.com, Deezer Research, Paris, France;, Idiap Research Institute, Martigny, Switzerland; Yashar Deldjoo, deldjooy@acm.org, Politecnico di Bari, Bari, Italy; Bruno Sguerra, research@deezer.com, Deezer Research, Paris, France; Markus Schedl, markus.schedl@jku.at, Johannes Kepler University Linz and Linz Institute of Technology, Linz, Austria; Manuel Moussallam, research@deezer.com, Deezer Research, Paris, France.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
Manuscript submitted to ACM

in general. For decades, progress in MRSs essentially meant improving performance on a series of retrieval-oriented subtasks, often evaluated offline, such as matrix completion, clustering, or sequence continuation. Soon enough, this framing revealed its limitations due to its markedly reductionist nature. It simplified the original question, *how do we design systems that make good recommendations?*, by sidestepping the harder one: *what actually constitutes a good recommendation?* Instead, the focus shifted to a more tractable problem: *how can we build systems that accurately predict what users will eventually consume?* While the two are arguably not the same, the latter offered a more straightforward way to evaluate the proposed systems in terms of algorithmic performance. And, as often happens in scientific practice, the measure gradually became the objective [24].

Attempts to bridge the gap between the evaluation paradigm and the original question have largely focused on introducing user studies that go beyond accuracy-based evaluation [93, 165], or on addressing important adjacent issues such as biases, fairness, and algorithmic accountability [42]. Nonetheless, the underlying paradigm has remained largely unchanged for decades, and RS papers continue to be evaluated mainly through accuracy metrics on standard benchmark datasets. Although critical perspectives have generally been welcomed within the community [32], they have rarely succeeded in fundamentally transforming this evaluation framework. Put simply, in the absence of a clear and practical alternative, the simplistic accuracy-driven paradigm has persisted.

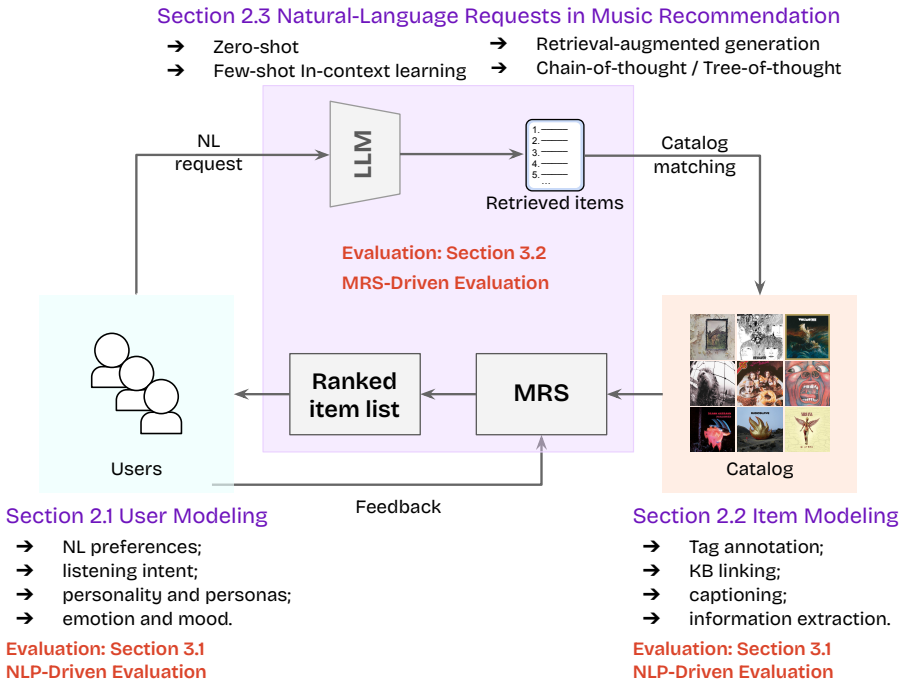
A key reason why this paradigm is particularly limiting in the music domain lies in the intrinsic nature of music consumption itself. Unlike many other recommendation domains, music listening is deeply contextual, affective, and embedded in everyday activities. The same user may seek entirely different music depending on their momentary goal, such as focusing while working, regulating their mood, socializing with friends, or expressing their identity. For instance, a user might request "calm instrumental music for studying," "energetic tracks for a workout," or "songs similar to this but less melancholic." These needs cannot be fully captured by past consumption alone, as they reflect transient intentions, subjective descriptions, and contextual rather than stable long-term preferences. As a result, the next item predicted from consumption patterns may not correspond to the music that best satisfies the user's current needs.

Music recommendation is further complicated by the nature of music items and the feedback they generate. Music is multimodal and the metadata used to describe music, such as genre, style, or mood, are often subjective, culturally dependent, and inconsistently defined. For example, descriptors such as "relaxing," "dark," or "summer vibe" may refer to different musical characteristics depending on the listener and context. At the same time, user feedback is mostly implicit, consisting of noisy behavioral signals such as play counts, skips, or listening duration, which provide only indirect and ambiguous evidence of preference. These properties make it difficult to rely on a single ground truth, to represent music and users precisely, and to evaluate recommendation quality using conventional metrics alone as they reflect challenging fundamental characteristics of music consumption instead of objective constraints. This stands in contrast to domains such as movies or e-commerce, where preferences can be more easily specified using factual attributes such as actors, or objective product features.

The emergence of Large Language Models (LLMs) and their rapid adoption in RSs disrupts this long-standing evaluation practices, as the fundamental assumptions of the established paradigm no longer hold [35, 37]. LLMs are generative models: they do not compute similarity scores nor rank items in the conventional sense. Therefore, they cannot be reliably evaluated using standard accuracy metrics. More critically, such metrics were not designed to account for hallucinations, knowledge cutoffs, the influence of training data sources, or personalization.

These limitations are particularly consequential in the music domain, where recommendations must align with subjective contextualized NL requests or descriptors provided by users, such as mood, activity, or cultural context, resolve ambiguous artist or track names, and remain grounded in large, rapidly evolving, and long-tailed catalogs.

Fig. 1. Paper's overview as a generic diagram presenting music recommendation with LLMs



Moreover, the very notion of train/test separation becomes largely irrelevant for black-box models such as ChatGPT—especially given that these models have likely been exposed to most of the public datasets commonly used for RSs evaluation [41], making reproducibility and fair comparison of results difficult, if not impossible.

At the same time, LLMs have inspired a new wave of methodologies and can even be applied directly to the core MRS objective, that is, users can simply ask an LLM for music recommendations, much as they would ask a human acquaintance, or they could ask LLMs to assess recommendations in a role as judges.

Despite the breadth of challenges that LLMs introduce to traditional MRSs evaluation, many works on LLM-based MRSs have continued to rely on standard accuracy-based metrics, often without fully considering what these numbers signify for models that generate text rather than predict or rank discrete items. Regardless of where one stands in the debate over the disruptive nature of LLMs, their growing use in MRSs represents a valuable opportunity to revisit how we evaluate RSs. It encourages the exploration of new evaluation perspectives that better align with the human experience of searching for and receiving music recommendations, while also prompting reflection on what constitutes a poor recommendation in generative settings, such as cases of hallucination or factual inconsistency. Effectively, it also requires taking inspiration from the field of Natural Language Processing (NLP) to evaluate the text produced by LLMs.

Against this background, we provide the article at hand, which constitutes a synergistic combination of survey and position paper, and is structured as follows: In Section 2, we provide a comprehensive overview of how LLMs have transformed various aspects of user modeling, item modeling, and recommendation, particularly in the music

domain. Then, in Section 3, we first discuss evaluation at the general level, by considering each scenario as a text generation task and exposing what is done in adjacent research domains. We also draw inspiration from existing work in music recommendation to outline current strategies for measuring both success and risk, while introducing a detailed set of evaluation dimensions and challenges that we argue should be considered in any study involving LLMs as recommenders. Figure 1 provides a detailed overview of the article's structure. The main contributions are:

- A detailed overview of background, challenges, and LLM-induced changes in MRSs, with an emphasis on the specificity of music consumption (Section 2).
- An NLP-inspired evaluation taxonomy that goes beyond traditional IR metrics and covers a wider range of scenarios, including human vs. automatic evaluation and reference-based vs. reference-free settings (Section 3.1).
- A concrete MRS-oriented evaluation framework with operational success metrics, failure-side diagnostics, and working examples for LLM-based music recommendation (Section 3.2)

2 LLMs in Music Recommendation: Domain-specific Challenges and Opportunities

In the following, we focus on three key aspects of MRSs where LLMs have been applied: user modeling (Section 2.1), item modeling (Section 2.2), and natural language-based music recommendation (Section 2.3). For each of these aspects, we structure the discussion around three main points: 1) background or fundamentals of the task; 2) music-specific challenges and domain considerations, highlighting what makes the problem unique in the context of music; 3) impact of LLMs, analyzing how generative language models can or have already changed methods and capabilities in the respective task.

Past research on music item or user modeling for MRSs has mainly focused on two directions: (1) enriching items/users with explicit, often high-level information (e.g., tag annotation or extraction, knowledge base linking) and (2) representing items/users as features for use in content-based or hybrid recommender systems (e.g., representation learning). In what follows, we concentrate on the first aspect, which directly leverages the generative capabilities of LLMs to enrich the item or user information. The second aspect, involving text embedding and feature representation, is left for future work, as it encompasses broader approaches that extend beyond LLMs to include other text encoders and multimodal representation models.

2.1 User Modeling

A first use case of LLMs in the context of MRSs is to create and extend user profiles in an accurate and comprehensive way, tailored to the music domain and recommendation task at hand. In the following, we first describe LLMs' potential for general music preference modeling, pertaining to traditional information such as consumption or interaction histories. Subsequently, we discuss their merits for advanced user modeling, involving personas, affective cues, listening intents, and cognitive effects.

Background. Most automatic MRSs rely on modeling user preferences. These preferences are often represented as vectors or embeddings, learned from interaction data using neural architectures in collaborative filtering settings [73, 79, 90]. A key limitation of such embeddings, derived from black-box models, is their opacity: they are not easily interpretable. This hinders both the generation of explanations for users and the investigation of the underlying decision processes. For user representations to be effective, a comprehensive collection of interaction data is required, linking each user to a diverse set of items. This requirement introduces several drawbacks. For instance, cold-start users, those who have just joined the platform, lack a relevant representation and thus receive poor recommendations. Similarly, if

a user wishes to steer their recommendations toward new interests, the system typically demands a substantial number of interactions with items from these new domains before the model can adapt.

User profiles can also span to the level of *personas*, i.e., groups of users that share common characteristics or tastes. These can be defined, among others, at the level of demographics (age, gender, place of residence), involve cultural aspects (religion, beliefs, norms), or socio-economic factors (wealth, income, social class). Together with models that connect one or more of these factors to music preferences, e.g. [106, 210], such personas have been used to overcome cold-start scenarios. For instance, the MRS can only use the user's belonging to a demographic group, to make recommendations, in a traditional demographic filtering approach. Alternatively, the definition of personas can also reflect more closely certain music-related patterns or dispositions. For instance, personas can be modeled as having different preferences for musical diversity, popularity, or mainstreaminess [12, 99].

Several works have provided evidence for a strong connection between music preferences and *psychological aspects*, including *personality* [123, 163], *listening functions* and *intent* [111, 166] as well as *affect* [58, 134, 139]. The purposes or reasons why people listen to music have been studied intensively in music psychology literature [111, 166], resulting in the emergence of three major trajectories, i.e., mood regulation, self-awareness, and social relatedness. *Personality*, often represented via the Big Five Model (aka Five Factor Model), describes traits along the dimensions of openness, conscientiousness, extraversion, agreeability, and neuroticism [121]. While personality can also be part of a persona description, it goes beyond common persona definitions, especially in the context of music because of its relationship with musical perception and preferences. For instance, correlations have been identified between personality traits and acoustical properties such as loudness or tempo [123], genre preferences [153], and perceived emotion [163]. Based on such insights, a few attempts to integrate personality information into the music recommendation process have been made [108, 112, 175]. Furthermore, modeling affective states of listeners is strongly related to modeling *listening intents*, as mood regulation is one of the dominant reasons why humans listen to music [166]. Eliciting affective cues such as *emotions* (short-lasting reactions in response to a stimulus), *mood* (longer-lasting episodes independent of a concrete stimulus), or *sentiment* (general positive or negative attitude towards a concrete item or item group) is thus important to tailor recommendations to the user's current affective state.

Music-specific challenges. Compared to domains such as books, movies, or e-commerce [64, 148, 223], music consumption presents a number of particular challenges for user modeling in the context of MRSs:

(1) Music catalogs are extremely large, often spanning tens of millions of tracks. Moreover, music catalogs tend to be very rich in terms of culture, genre, year of production etc. As a result, users interact with hundreds or even thousands of unique tracks over time [165].
(2) On music streaming platforms, the most common form of feedback is implicit. Implicit feedback refers to user actions and behaviors that are not explicitly provided as signals of preference. In this setting, feedback is typically derived from listening time or from binary listening events defined by a time threshold, often set at 30 seconds. The challenge stems from implicit feedback being a noisier signal of interest than explicit ratings or reviews [173]. Moreover, in the absence of explicit textual feedback, user preference modeling must rely on item metadata, whose quality varies considerably [120].
(3) User musical preferences are complex, often spanning both short- and long-term dimensions [170]. Short-term preferences reflect recent interactions and can be quite noisy, while long-term preferences capture more stable tastes developed over years. However, long-term preferences may also become outdated and lack novelty. Moreover, musical preferences are highly contextual. For example, soundtracking, selecting music for a given

activity, is a very common way of consuming music [56, 118]. In different contexts, users will change the type of music they consume to match the mood or the energy pertaining to their current activity: someone will play energetic music when working out or calm music for studying. This can drastically change the user's preferences, something that a static user representation might not be able to pick up.
More broadly, soundtracking illustrates a functional use of music, where listening is guided by an explicit goal rather than by genre or artist affinity. In many everyday scenarios, music is used to support an intended outcome (e.g., focusing, regulating mood, or increasing motivation). These functional aspects of listening are consistent with evidence that music can influence human cognition, emotion, and behavior [152]. Thus, music consumption is often highly situational, shaped not only by enduring taste but also by immediate purpose and context.

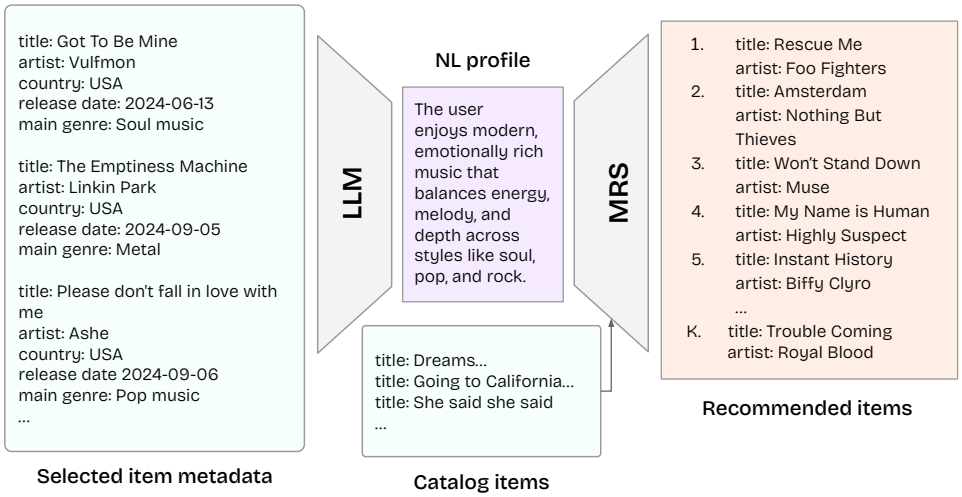
(4) Affective responses can be quite subjective in nature when it comes to music. So are their impact on music preferences. A key distinction in this regard is between perceived and induced emotions. Music can both convey an emotion and evoke an emotional response, and the gap between these two aspects is non-trivial. Perceived emotion refers to the emotion recognized as being expressed by the music (e.g., identifying a piece as "sad"), whereas induced emotion corresponds to what the listener actually feels while listening, which can vary substantially across individuals [177]. This inconsistency further complicates affect modeling, particularly in the absence of explicit user-provided affective feedback.
An additional factor making the use of affective cues challenging is that music is a highly multimodal artifact [38] and different channels of delivery (e.g., audio, lyrics, album cover, video clip) may elicit different emotions. Moreover, emotions may dynamically evolve over the time a user listens to a music piece. How to acquire and integrate such emotional trajectories in the recommendation process is still a complex problem.

Key changes introduced by LLMs. With the introduction of LLMs, a new alternative to traditional opaque user embeddings has become viable: the use of natural language representations of user preferences [64, 148, 156, 170, 223]. In this setting, LLMs act as summarization devices, producing high-level representations of users' preferences from collections of interacted items. The use of textual taste representations offers many advantages, particularly their interpretability. Because these profiles are understandable, users can see what the system "knows" about them, increasing transparency and building trust. Moreover, in cold-start situations, users can explicitly state their preferences, enabling the system to generate recommendations. Furthermore, if users are allowed to edit their profiles, they could take control of their recommendations by steering them toward new interests and away from content they no longer find relevant.

A common method for deriving natural language preferences with LLMs is to concatenate explicit feedback, such as item reviews, and metadata from consumed items into prompts and use the LLM to summarize this information into a profile [64, 148, 170, 223]. It is widely acknowledged that such profiles need to be concise in order to be scrutable. In this process, LLMs aggregate the most relevant information encoded in users' consumption histories. In a second step, after profiles are generated or collected, recommendation can again proceed in multiple ways: either by projecting both user profiles and item metadata into a shared latent space [64, 170] or by employing a second LLM, employed here as recommender (described in Section 2.3). A representation of this method is shown in Figure 2.

Given that LLMs are nowadays often perceived to exhibit human-like responses and show traces of "thinking" or reasoning, it is natural to investigate, and potentially leverage, the *psychological traits* they may simulate, adapting concepts and mechanisms from human psychology such as personality, emotions, or cognitive effects and biases. LLMs could enable new ways of approaching personality-related tasks in the domain of music recommendation. For instance, they can be used to create personality profiles of their users (e.g., by predicting traits from the user's prompting

Fig. 2. Representation of the method to derive NL user preference profiles from consumption data.



history or by carrying out a conversation with the user guided towards inferring their personality using relevant personality assessment questionnaires) [180, 213]. Enriched with correlation models between personality and music traits or preferences such as those proposed in [112, 123], LLMs could then further tailor recommendations accordingly. Also, LLMs appear to infer persona-related aspects such as musical diversity, popularity, or mainstreaminess [12, 99] from the user's listening history with some success. Research on persona-based music recommendation can draw from recent large datasets of LLM-generated personas, e.g. [67], that provide a synthetic dataset of 1 billion personas. While currently not tailored to the music domain, corresponding research can inform the creation of music-specific persona datasets, which in turn could be used to (1) support the classification of users into personas from their conversation with the LLM and (2) facilitate the simulation of music-related conversations with the system.

Regarding listening intents, LLMs provide a great tool to consolidate and extend these findings from music psychology. For instance, integrating a data-driven approach with the results of the empirical questionnaire-based study from [166], Hausberger et al. leverage LLMs to augment and cluster 129 listening functions into 32 broader intents and provide a simple interface that enables music search by intent [72]. Another pathway could be to infer the listening intent of the users from their conversations with an LLM, either through additional classifiers or directly through LLM-guided questions. Ultimately, the listening intent could be used by traditional recommendation approaches, integrating it as contextual factor, or again directly by the LLM.

Surprisingly little research has been devoted to using LLMs for finding music that fits a specific *emotion*. Nevertheless, LLMs have shown some potential to identify emotions from conversations [55, 62]. Linking these to individual music preferences and using the resulting user model to tailor music recommendations, however, is still an open challenge with only very recent and limited research available [192, 194]. This is a highly valuable direction given the importance of emotion and mood in music perception and use. It is important to note that emotion and mood are different concepts from a psychological perspective, but often used interchangeably in the more technical RSs research works. Taking a psychological view, emotion refers to an affective response to a stimulus that is short-lasting but highly intense, while mood refers to a longer-lasting affective state of lower intensity that is not necessarily the result of a (single) stimulus.

Table 1. Extractive tasks for data augmentation when richer descriptions of music items are available.

Extractive Task	Description
Tag Extraction	Extract attributes or tags of music items from text (e.g., music genres, lyrical themes, instruments or voice characteristics)
Named Entity Recognition (NER) *Coreference Resolution*	Detect names or mentions of music items (e.g., artists, albums, or tracks) from text. In creative domains like music, challenges stem from the frequent use of common words to denote named entities, and input text being often very noisy (e.g., user natural language queries or transcribed voice commands).
Sentiment or Negation Identification	Identify music items that are discussed with negative sentiment or explicitly disliked, which can be valuable for personalization and content filtering.

2.2 Item Modeling

While collaborative filtering RSs primarily rely on user–item and item co-occurrence in streaming sessions, content-based or hybrid MRSs focus on a deeper understanding of an *item's features* during the recommendation process. These features align better with how humans perceive music, offering more support for explainable recommendations, enabling discovery, and item cold-start situations [38, 164].

Background. Item features relevant to MRSs range from low-level attributes extracted from audio or lyrics to higher-level features such as user-generated content (e.g., playlist titles or music tags). Deldjoo et al. [38] propose an "onion model" to categorize musical content relevant for recommendation, placing *audio* at the core, followed by *embedded metadata* from music providers (e.g., title, artist, album, lyrics, label, recording year), and then outer layers of increasingly expressive, subjective, or culturally influenced features that reflect personal or community perceptions of music. In particular, *expert-generated metadata* often focuses on musical aspects such as genre, style, or mood, but may also include details about music properties (e.g., instrumentation, vocal attributes, lyrical themes) or artist information like biographies. *User-generated metadata* includes item descriptions such as tags, lyric interpretations, and music reviews, typically created on social media or forums (e.g., last.fm[1], reddit[2], SongMeanings[3]) or within music streaming platforms when naming playlists or interacting with tracks after a search.

Embedded metadata and editorial information, such as genres, are typically available when music is ingested into streaming services. However, this data could be incomplete, ambiguous, or inaccurate. To tackle this, past works have explored matching musical items with knowledge bases or editorial encyclopedias, as well as aligning existing musical taxonomies or graphs to improve metadata verification, completeness, and music contextualization (e.g., adjusting music tagging taxonomies to a specific country) [48, 49].

Another approach to describing music, e.g., by its moods, styles, and instrumentation, is to rely directly on the audio signal. Many works in the Music Information Retrieval (MIR) community focus on tasks such as audio auto-tagging [83] or automatic audio classification [147] (e.g., by genre, mood, or listening context); voice and instrument identification [150]; lyrics transcription [225]; and music captioning [60, 66].

When richer textual data related to music, such as reviews or music/artist descriptions, is available, established NLP techniques could be used for aspect extraction [14], named entity recognition [47], and sentiment or negation identification [154], as detailed in Table 1.

[1] https://www.last.fm
[2] https://www.reddit.com/r/rock/
[3] https://songmeanings.com

Music-specific challenges. Unlike domains such as books, news, or e-commerce recommendations, music content is largely non-textual apart from lyrics. However, lyrics are copyrighted, making their use difficult in a setting of reproducible research, as experimental datasets cannot be shared [109, 129].

In the context of music streaming services, textual information is often limited to short forms [48, 49], and common NLP techniques struggle with these short or noisy texts [5] (e.g., user tags or playlist titles, transcribed user queries). When richer textual data is available, such as music reviews or social media posts, the text nature of track titles and artist names introduces additional challenges [47]. Their vocabulary often overlaps with tagging terms. For example, "love" might refer to a track title, a mood, or a lyrical theme. This ambiguity complicates key tasks such as named entity recognition and artist disambiguation (linking entities to the catalog or knowledge base). Also, detecting negations as a way of expressing dislike (e.g., disliked music artists or genres) remains a significant challenge, even for LLMs [65, 186].

When relying on audio as input, music auto-tagging was until recently limited to fixed vocabularies, performing well for some tags but failing with contextualized or user-specific ones (e.g., music for "travel", "running", or "relaxing") [83]. Similarly, early audio captioning works, prior to transformer-based models, often produced unnatural or even grammatically flawed descriptions [45]. Although newer models have improved linguistic fluency [60, 66], some studies show that audio is not always effectively leveraged as a modality [198].

Finally, music, as a cultural product, is perceived differently across communities [22, 49]. The same artist may receive different genre tags depending on the annotation practices of music providers or listeners, or the localization of Wikipedia editors. This variability is even greater for mood- or context-related tags, where personalization is essential but often absent (e.g., "party music" can mean very different things to different users [122]). Consequently, aggregating and reconciling diverse knowledge sources for tag annotation, or music description generation in general, remains a significant challenge.

Key changes introduced by LLMs. LLMs provide powerful capabilities for augmenting item information and have been leveraged in several complementary roles, each addressing different data-augmentation needs, as follows:

LLM as Annotator or Knowledge Base. Trained on vast amounts of online text, LLMs have become alternatives to knowledge bases or human annotators, even in zero-shot settings [9, 182, 216]. Thus, LLMs could generate user-friendly tags or keywords that describe a track, including its music genre or style. LLMs could also be used to suggest editorial metadata such as release dates, country of origin, or artist details, when these are missing. Fine-tuned LLMs specialized in the music domain can go further by producing rich, descriptive content about items [46]. Compared to other knowledge bases, LLMs can offer a degree of personalization in generating music metadata or descriptions, whether through prompt manipulation (e.g., including characteristics of the target user or community), in solutions based on Retrieval-Augmented Generation (RAG) that interface with multiple tag taxonomies, or when used as personalized AI agents [30, 115].

LLM as Information Extractor. LLMs have proven highly effective at extracting specific information from unstructured text without even requiring a domain-adaptation or fine-tuning phase [71, 159]. For example, given album reviews (e.g., Pitchfork[4]), social media posts (e.g., musicsuggestions subreddit[5]), or editorial content (e.g., rateyourmusic[6]), LLMs could identify and extract mentions of items or specific aspects associated with them. Moreover, LLMs could be directly integrated within conversational RSs and extract examples of liked and disliked music as well as preferred and disliked musical attributes [208] (e.g., "I like classical music but dislike repetitive vocals"). This process involves

[4] https://pitchfork.com
[5] https://www.reddit.com/r/musicsuggestions
[6] https://rateyourmusic.com

leveraging LLMs for Named Entity Recognition (NER), aligning an open vocabulary of short phrases with a fixed taxonomy, and high-level text reasoning to interpret expressions of dislike.

LLM as Summarizer or Captioner. LLMs are highly effective in generative tasks compared to any preceding models, particularly when the goal is to transform input data into new textual outputs with different structures or distributions from the source. In this role, LLMs are known to perform tasks such as: 1) *Summarization*—summarizing multiple sources of music item information into a concise, coherent summary or description; for instance, generating a short item caption from longer user descriptions focused on the interpretation of lyrics [128]. 2) *Item co-occurrence summarization*—analyzing patterns of music item co-occurrences and their metadata, then translating these insights into descriptive text that highlights relationships between items; an interesting application is generating smooth transitions or spoken segues between two tracks, similar to the commentary or explanations heard in radio shows [59]. 3) *Music captioning*–creating new content such as item captions or profiles, similar to the user profiles [43, 182].

While all these roles and examples focus on textual data as input, *Multi-Modal Language Models* (MMLMs) [39, 94], which can incorporate other signals such as an item's audio, can also be leveraged to augment music items with additional information such as audio characteristics or album cover artwork.

2.3 Natural Language Requests in Music Recommendation

We further focus on the setting where users express their music requests and preferences in natural language (NL). While communicating music preferences through language is arguably the most natural and socially grounded way of requesting music, resembling how people exchange recommendations among friends [208], it remains a highly challenging scenario for modern Music Streaming Platforms (MSPs). Thanks to their strong textual understanding capabilities, LLMs have made this one of the most affected, and perhaps most promising application areas.

Background. Given how natural it is for humans to articulate their needs in NL, MRSs with NL interfaces have emerged as a promising way for users to interact with the catalog [87]. While most MSPs provide a search bar that allows users to make textual queries to access the catalog directly, their functionalities remain limited. The search bar represents a high-control interaction mechanism, enabling users to access specific content, for instance, music from a particular artist or album [169]. Users also rely on search when they only have a "seed of an idea" about what they are looking for [101], such as "music to work out" or "music for studying." Despite the inherent limitations of traditional keyword-based search, users often adapt to its functioning and constraints over time.

The introduction of LLMs makes it possible to move beyond keyword matching or predefined frames in the case of traditional voice interaction [87], and to support more complex and expressive user requests. Several studies have explored integrating them into recommendation interfaces that connect users' NL requests to the catalog [33, 122, 132]. In this context, LLMs could draw on their extensive general "knowledge" of music to identify or describe relevant items. Trained on massive and diverse sources of texts, predominantly from the web, they are able to capture and express musical attributes, such as tempo, rhythm, and timbre, through linguistic associations [21]. This opens the possibility of bypassing traditional collaborative filtering pipelines and potentially mitigating issues such as popularity and exposure biases, which arise when models are trained solely on logged interaction data [25, 81, 120, 141, 188]. However, LLMs also introduce a new set of challenges that must be carefully addressed.

Music-specific challenges. Music makes for a particularly challenging recommendation scenario. It is a rich and multifaceted medium that spans cultures, countries, decades, and genres [38]. As also discussed in Section 2.1, music catalogs typically contain millions of tracks, each of which can be described through metadata such as release year, genre,

or country of origin [165]. In addition, the audio signal itself carries a wealth of information that can be extracted—such as mood, tempo, or instrumentation [130, 138]. Music perception is both cultural [97] and dynamic [171, 172], varying across individuals and communities. Moreover, music consumption is often highly contextual: as music influences human cognition, emotion, and behavior, it can be deliberately employed for specific goals such as improving concentration, maintaining motivation, or expressing social identity [152]. Consequently, users often listen to music as a tool for emotional regulation, or selecting tracks that match a particular context or activity [118].

These diverse dimensions make music an especially complex domain for NL requests. Users may ask for highly specific content (e.g., "Beatles songs from their early years") or for abstract, affective experiences (e.g., "chill music with intense energy"). The challenge lies in bridging these linguistic expressions with the underlying catalog representations—while also accounting for users' personal preferences, cultural background, and situational context.

Adding to this difficulty, new music tracks are continuously uploaded to the catalogs used by MSPs, creating persistent cold-start issues [52]. When these new tracks are absent from the LLM's training data, the model tends to favor older or more popular tracks, systematically overlooking newly released material. This is particularly problematic since many users rely on MRSs precisely to discover new music [6]. Ultimately, NL-based music recommendation requires systems that can interpret language, not only semantically, but also affectively and contextually, aligning textual intent with both the structure of the catalog and the lived experience of music consumption.

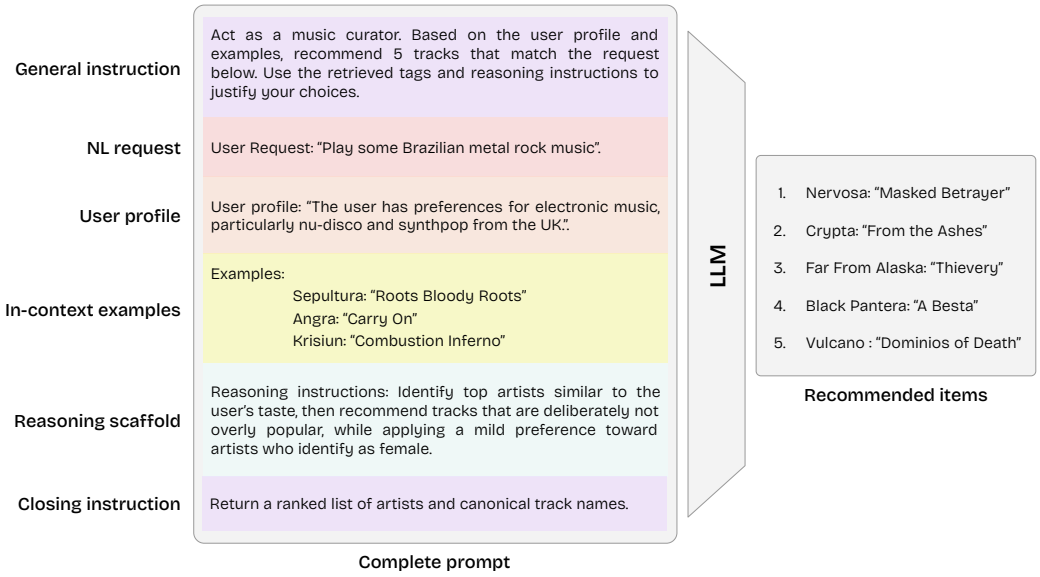
Key changes introduced by LLMs. LLMs enable more flexible and expressive user queries in music recommendation. Moreover, through prompt design, their behavior could be guided or constrained, for instance, to mitigate biases related to popularity, gender, or other social factors [34]. In the following, we detail the different ways LLMs have been employed as music recommenders. We focus on the scenario in which a **ranked list of musical tracks** is produced in response to a user's NL request. This setting closely aligns with traditional top-K recommendation and playlist-generation tasks [33]. We define the following task components:

- *Music catalog:* track identifiers, canonical names, artist/album relations, and editorial or community metadata. Catalogs often contain multiple tracks sharing the same canonical names (e.g., homonymous artists, different versions of the same song across collections) and exhibit noisy or heterogeneous metadata, often supplied by labels, producers, or users;
- *Resolver/knowledge base:* used for entity validation and disambiguation (e.g., MusicBrainz or the Music Ontology are often used for schema-level reasoning) [125, 146];
- *NL user request:* e.g., "play some obscure bolero guitar pieces," "recommend me melancholic violin cues from anime soundtracks," or "play Brazilian forró with accordion solos.";
- *NL user profile:* long- and short-term music preferences expressed in natural language [170];
- *In-context examples:* few-shot seeds / examples given as a list of tracks, artists, or other musical entities;
- *Retrieved documents for RAG:* e.g., metadata snippets, tags, reviews, or new-release notices [100];
- *Reasoning scaffold:* e.g., Chain-of-Thought (CoT) or Tree-of-Thought (ToT) prompting [199, 203].

As outlined by the list of task components, we assume that generation is grounded in a *resolver* or *knowledge base* (KB) and may be further structured through prompting scaffolds. Specifically, as the MIR community has long emphasized domain-specific challenges such as *entity disambiguation* (artist, release, recording) [124, 149], *heterogeneous metadata* [16], and *long-tail discovery* [23, 52], we explicitly separate the catalog, which is often noisy, from KBs, such as MusicBrainz[7],

[7] https://www.musicbrainz.org

Fig. 3. Prompt example with the different task components highlighted.



Discogs[8], or AcousticBrainz[9]. The KB is then used to *validate the existence of and disambiguate entities* (e.g., homonyms, release variants, remasters), and to map generated item names back to the catalog in order to filter out hallucinations. Furthermore, a prompting operator composes these elements into a single prompt and a post-processor deduplicates, validates, and re-ranks the resulting list of tracks. Figure 3 illustrates a complete prompt, split into panes: different recommendation tasks correspond to combinations of these elements. Next, we describe them in detail.

Zero-Shot Recommendation (input: user NL request-only, possibly conditioned by user profile). In this setting, following the example in Figure 3, only the user request (e.g., "Play some Brazilian metal rock music") is required, optionally accompanied by the NL user profile.

While classical RSs rely heavily on collected user feedback, zero-shot recommendation reframes the task as natural-language understanding, where the LLM maps open-vocabulary constraints to catalog items [17]. In the absence of a user profile, the model relies entirely on its linguistic priors. LLMs internalize broad cultural knowledge and can parse open-vocabulary descriptions, including compositional constraints, say, in terms of affect (valence, arousal, and dominance) [196], key/mode, or content control (e.g., "avoid harsh vocals").

In profile-conditioned zero-shot recommendation, personalization is achieved by prepending a readable description of the user's taste. The profile is interpretable, editable, and directly leverages behavior, unlike opaque latent vectors [200]. This "profile-as-text" approach enables controllable personalization without modifying model weights, allowing users or curators to make immediate corrections [200].

[8] https://www.discogs.com
[9] https://www.acousticbrainz.org

Zero-shot recommendation is particularly valuable in *exploratory* and *cold-start* scenarios as it operates without requiring any historical user interactions, instead treating the recommendation task as a language understanding and generation problem [17, 200]. Recent research shows that zero-shot LLM-based recommenders are promising in handling long-tail items and under-described genres that lack sufficient behavioral data [74, 78]. However, the zero-shot approach carries inherent risks: without explicit grounding in retrieved documents or contextual information, these systems are susceptible to popularity bias and may generate plausible but incorrect recommendations (i.e., hallucinations) [88].

The fundamental challenge stems from the model's reliance on parametric knowledge, i.e., information encoded in its weights during pre-training. Recommendation quality thus depends critically on two factors: accurate entity resolution that correctly maps user queries to items in the KB, and robust post-processing to validate outputs. When the model's parametric knowledge becomes stale relative to catalog updates, the zero-shot setting becomes particularly vulnerable to hallucinations, especially for niche or newly released items that were not present during training [88].

Despite these limitations, zero-shot approaches enable innovative user experience patterns. For example, systems can implement editable "taste cards" that allow users to modify their preferences and immediately see updated rankings without requiring model retraining [64, 131]. To harness this flexibility while mitigating risks, practitioners should implement stricter entity resolution mechanisms and post-validation procedures to prevent the system from generating confident but incorrect recommendations [88, 100]. In summary, zero-shot recommendation is a powerful tool for generating diverse suggestions and supporting exploratory discovery, provided that evaluation frameworks prioritize grounding accuracy and maintain balanced exposure across the item catalog.

Few-Shot In-Context Learning (ICL) (input: user request and in-context examples, possibly conditioned by user profile). The zero-shot approach can be extended by adding in-context examples, short listening histories or seed items, that capture user preferences. These examples can be sourced in two primary ways: from the target user's own interaction history (*personalized few-shot*) [34] or from similar users' profiles in a collaborative setting (*collaborative few-shot*) [11, 205, 217]. Collaborative approaches are particularly valuable for cold-start scenarios or enriching sparse user profiles.

Recent work explicitly designs systems to select similar user histories as in-context examples, enabling LLMs to actively participate in choosing these high-quality demonstration examples [217]. For instance, given the example in Figure 3, the items "Sepultura - Roots Bloody Roots", "Angra - Carry On", and "Krisiun - Combustion Inferno" serve as demonstrations to nudge the generation toward similar curation. Whether drawn from personal history or collaborative neighborhoods, the model leverages a few explicitly liked tracks or micro-histories to enable immediate personalization, well-suited for adapting to new or shifting tastes without re-training [11]. Few-shot ICL is particularly useful in session-based recommendation, where short-term intent can be efficiently reflected through a curated list of seed examples (see, for example, AdaptRec [217]).

However, selecting appropriate seed items requires careful consideration regardless of their source. Consumption logs are multi-intent and dominated by implicit signals (skips, repeats). If seeds are sampled naively from the chronological tail, recent interactions may overshadow long-term preferences and inject noise [170]. In collaborative settings, this challenge intensifies: selecting exemplar users whose preferences are genuinely similar requires careful similarity computation to avoid amplifying biases from noisy neighborhoods [217]. Cross-lingual aliasing and transliteration are also a common issue: if all seeds belong to a single language, the model may amplify that linguistic or regional bias unless the selection is balanced [49]. Moreover, research on session-based recommendation documents strong order sensitivity. ICL inherits this hazard unless the selected examples are stress-tested with shuffle, drop, or paraphrase perturbations [113, 143].

Beyond personalization, ICL also opens new design levers for controllability. LLMs can interpret heterogeneous examples, such as track IDs, textual notes (e.g., "no harsh vocals"), or micro-histories, and synthesize them with open-vocabulary constraints, effectively producing a local curation policy "in context" [17]. Designers can encode fairness or exploration directly (e.g., balancing genders or regions in exemplar artists), a controllability lever that classical embedding-only collaborative filtering does not offer [200]. Because seed examples exist exclusively as textual tokens within the prompt context, practitioners can perform A/B testing on prompt templates, system roles [34], and instructions without requiring model retraining or parameter updates [17].

Retrieval-Augmented Generation (RAG) (input: user request and retrieved documents, possibly conditioned by user profile). RAG enriches the prompt with retrieved documents drawn from either the catalog or an external KB (e.g., item metadata records, artist profiles, release information, user reviews, social tags, or new-release announcements) [100]. Evidence retrieved from the catalog enables the model to cite specific release groups and track identifiers, which can subsequently be mapped back to the catalog for execution. The fundamental goal is to ground recommendations in factual evidence and mitigate hallucinations by combining parametric knowledge encoded in model weights with non-parametric access to up-to-date external information [63]. This paradigm is increasingly adopted to ensure catalog freshness by leveraging current information from external sources, since new releases and re-releases appear continuously in music streaming platforms [46].

Music catalogs present unique retrieval challenges that distinguish them from general-domain applications [162]. The same song may exist in multiple versions—original recordings, remastered editions, live performances—and metadata is frequently inconsistent or multilingual across databases. To address such complexities, RAG systems might employ two complementary search strategies. First, exact matching through catalog identifiers and structural relationships efficiently retrieves known items with high precision. Second, semantic search, which compares the contextual meaning of text rather than surface-level keyword matching, discovers relevant music even when users describe requests informally or in other languages [201]. The system then validates that retrieved items exist in the catalog and resolves ambiguities, such as multiple artists sharing identical names or cover versions among different recordings.

Two practical considerations fundamentally shape RAG implementation in MRSs. First, computational and economic efficiency: including excessively from retrieved content in the prompt increases both token consumption and processing latency, requiring careful ranking and selection of only the most relevant information to maintain real-time response guarantees. Second, catalog freshness and dynamic updates: as retrieval indices reside external to the LLM's parametric weights, music streaming services can immediately add newly released tracks to search indices without model retraining or fine-tuning [100]. This decoupled architecture enables LLMs to effectively bridge informal user requests (e.g., "upbeat workout music") with formal catalog metadata (genre classifications, tempo measurements, energy ratings) [46]. When appropriately validated through entity linking and relevance filtering, this approach enables scalable, factually grounded recommendations across millions of tracks while maintaining responsiveness to catalog evolution [63].

Chain-of-Thought / Tree-of-Thought (CoT/ToT) (input: user request and reasoning scaffold, possibly conditioned by user profile). Reasoning scaffolds guide LLMs to decompose complex recommendation tasks into explicit intermediate steps. For example, a CoT prompt for creating a workout playlist might instruct: *"Generate a 30-minute workout playlist by: (1) selecting high-energy tracks with BPM 140-160, (2) filtering for motivational lyrics, (3) ordering by increasing intensity, (4) ensuring smooth key transitions between adjacent tracks, and (5) justifying each transition."* Rather than producing recommendations through a single opaque generation pass, the model articulates each decision stage, making the curation process transparent and auditable.

CoT structures reasoning as a linear sequence of decisions (e.g., "filter by mood → constrain by era → match key/BPM → order for flow → justify transitions"), where each step's output feeds into subsequent stages [199]. ToT extends this by exploring multiple reasoning branches in parallel and aggregating across paths through self-consistency mechanisms or voting, improving robustness when single-path reasoning proves brittle [195, 203]. Both approaches enable the model to reason explicitly over multiple constraints, candidate selections, ordering criteria, and transition justifications.

Applied to music recommendation, these scaffolds provide three key benefits. First, *interpretability*: users and curators can inspect why each track was selected and verify reasoning against catalog metadata. Second, *controllability*: policy constraints (e.g., regional quotas, content restrictions, fairness objectives) can be encoded directly into the reasoning instructions without model retraining. Third, *multi-constraint handling*: the system can balance competing musical dimensions (mood, tempo, key, energy progression) through explicit deliberation rather than implicit optimization.

However, music-specific challenges require careful handling. Subjective tags, such as mood or energy, are culturally dependent and annotator-specific, and without explicit grounding in catalog data, models may construct plausible-sounding reasoning chains that rely on invented micro-genres, spurious correlations, or biases [85]. Even objective metadata such as key or instrumentation can vary across different recordings or remasters of the same composition. Then, effective playlist generation requires global sequence-level coherence (i.e., smooth transitions, narrative arc, energy flow) beyond individual track quality, necessitating evaluation at pairwise and playlist-level granularity [162].

Critically, reasoning scaffolds should remain decision-useful, improving selection quality and constraint adherence rather than producing verbose but uninformative justifications. When properly grounded in catalog data and validated against factual metadata, CoT/ToT enables scalable, interpretable music curation that supports complex multi-stage reasoning while maintaining transparency and controllability.

2.4 Discussion

We provided a detailed analysis of the three main areas of MRSs where LLMs have been applied: item modeling, user modeling, and natural language-based music recommendation. For each area, we structured the discussion around how the task has traditionally been addressed, the challenges arising from the specific characteristics of the music domain, and the impact of LLMs in addressing these challenges and existing methods and capabilities.

We now take a step back and summarize the key music domain characteristics and their implied challenges for MRSs in Table 3. The table is organized into three parts corresponding to the three areas above. However, several challenges cut across multiple areas: for example, the extreme size of music catalogs and the continuous influx of new content affect item modeling, user modeling, and natural language-based recommendation alike. The presented domain-specific particularities highlight that music recommendation involves more than matching users to items: it requires handling subjective, contextual, multimodal, culturally grounded, and highly dynamic preferences, often expressed through ambiguous or underspecified natural language that needs to be mapped on an extremely large and noisy catalog.

The exposed particularities and their corresponding challenges specific to the music domain should not be viewed in isolation, as they are interconnected and often reinforce one another. For example, the affective and context-dependent nature of music consumption interacts with the multimodal and culturally grounded representations of music, making both user modeling and item modeling inherently situational and subjective. Similarly, limitations in metadata quality and textual availability compound the semantic gap between natural language requests and catalog entities, especially in large, long-tailed catalogs where many items lack reliable tags. Implicit feedback and evolving short- and long-term preferences make it difficult to disentangle stable tastes from contextual or exploratory behavior, further complicating

Table 2. Particularities specific to the music domain and their corresponding challenges for MRSs. Challenges are highly interdependent, with user dynamics, item representations, and natural language interaction mutually influencing modeling and evaluation.

	Music-Domain Particularities	Implied Challenges for MRSs	Related Particularities
User-centric modeling and dynamics			
1.1	Evolving short- and long-term preference dynamics	New users provide little or no history, and their initial preferences may reflect transient exploration or highly specific, short-term interests rather than stable tastes. Preference modeling should account for noisy short-term exploration signals while balancing them with more stable, though potentially outdated, long-term preferences.	1.2, 1.3, 3.1, 3.2
1.2	Strong context dependence (music listening during an activity, for mood regulation, as a self-identity expression)	Frequent preference shifts demand context-aware and adaptive recommendation models; stable preference modeling is not sufficient; item modeling is tightly linked to the user's perception and context.	1.3, 2.1, 2.6
1.3	Highly affective and emotion-driven user demands	Subjective and dynamic responses demand highly personalized user/item modeling and evaluation.	1.2, 2.2, 3.4
Item representation and content characteristics			
2.1	Multimodal nature (audio, lyrics, metadata, cultural annotations, video)	Effective recommendation requires the integration of heterogeneous music facets, with some becoming central depending on the user's listening context, goal, or affect.	3.5, 2.3
2.2	Cultural and linguistic variability, subjectivity and personalization of music descriptors	Genre labels and music descriptors vary across cultures, regions, and communities, complicating personalization, contextualization, and fairness. Contextual tags such as "party" or "relaxing" are highly subjective and differ significantly across users.	1.3, 2.6, 3.4
2.3	Incomplete and conflicting metadata; limited expressiveness of audio-derived descriptors	Music annotations originate from multiple sources with differing standards, leading to inconsistencies and conflicts. Editorial metadata is often static and missing for a large part of the music catalog, limiting its reliability for representing items and matching them to users. Auto-tagging systems have historically relied on fixed vocabularies, failing to capture user-specific concepts.	3.1, 3.5, 2.2, 1.1
2.4	Limited access to music textual content	Music is primarily non-textual, and lyrics are often unavailable due to copyright restrictions, limiting accessible and reproducible datasets including lyrics.	2.1, 2.3, 3.5
2.5	Short and noisy textual signals	Available music-related text (e.g., playlist titles, tags, queries) is often brief, sparse, or noisy, limiting its descriptive power to represent music or users.	2.2, 2.6, 1.1
2.6	Music entity ambiguity in natural language queries, rare explicit expression of negation	Track and artist names often overlap with common words (e.g., "Love") or multiple music entities could share the same canonical name, complicating entity recognition and accurate linkage to catalog items. User dislikes are often expressed implicitly or through negation in natural language, which is difficult to detect reliably.	2.1, 2.5, 3.5
NL interaction and MRS evaluation			
3.1	Extremely large and long-tailed catalogs, continuous influx of new content	Persistent item cold-start, strong popularity bias, and difficulty ensuring fair exposure for less-known or novel music.	2.3, 2.4
3.2	Implicit feedback (play counts, skips, listening duration); passive consumption	Noisy preference signals that complicate user preference modeling and reduce the reliability of offline evaluation.	1.1, 3.4
3.3	Sequential, session- or playlist-based consumption	Quality goes beyond item-level relevance, depending on collection ordering, transitions, and coherence.	1.2, 1.3
3.4	Lack of clear ground truth for "good" music recommendations	Evaluation must extend beyond accuracy to discovery, diversity, and user experience; it is highly linked to the user's music listening goal, context, or affect.	2.2, 3.2
3.5	Complex and multi-dimensional natural language interactions to varied music facets	User requests range from precise factual constraints (e.g., specific artists or periods) to abstract, affective, or contextual descriptions (e.g., mood, energy, activity). MRSs need to interpret complex, open intent and map subjective language to multimodal and multi-cultural catalog entities.	2.1, 2.2, 2.3, 1.2, 1.3

both personalization and evaluation. These interrelations highlight that many core challenges in music recommendation emerge from the interaction between these areas, rather than from any single factor. This also explains why addressing one aspect (e.g., improving item representations) often has direct implications for other components.

Concluding remarks on LLMs for User Modeling in the Music Domain. LLMs introduce new ways to model users by enabling the construction of textual profiles from listening histories, onboarding selections, or conversational interactions, which is particularly valuable in the music domain where explicit feedback is scarce and cold-start is frequent. However, unlike domains such as movies or e-commerce, music preferences are highly contextual, affective, and dynamic, meaning that a single static profile is often insufficient to capture the diversity of a user's taste. As a

result, LLM-based solutions for music must go beyond generic profile summarization and support context-dependent representations, for example through multiple profiles or conversationally inferred personas that reflect different listening situations, emotions, or intents.

This illustrates that LLM-based approaches must be adapted to the specific structure of music consumption, where preferences shift frequently and depend on subjective and situational factors. At the same time, general LLM-based personalization methods do not directly transfer to music because they rely heavily on textual signals that are often sparse, ambiguous, or noisy in this domain. Profile construction from prompts or metadata may amplify noise, overlook negative preferences, or inherit inaccuracies from incomplete or conflicting music descriptors, while the affective nature of music requires interpreting nuanced emotional language without oversimplification. Moreover, LLMs must balance short-term listening signals with long-term preferences to avoid overfitting to transient behaviors. These domain-specific constraints imply that LLM-based user modeling for music requires additional mechanisms for grounding, calibration, and context awareness, as well as adapted evaluation protocols, compared to more text-rich and stable recommendation domains. Looking beyond these prompt-based configurations, recent work is increasingly framing recommendation as an agentic process, involving autonomous information seeking and, in some cases, coordination among multiple agents, thereby extending LLM-based recommenders from single-step generation toward more deliberative and interactive decision-making [10, 105, 117].

Concluding remarks on LLMs for Item Modeling in the Music Domain. LLMs have expanded item modeling in music by acting as annotators, information extractors, and summarizers, enabling the automatic generation of tags, metadata, and natural language descriptions from diverse sources such as reviews, social media, or editorial metadata. These capabilities are particularly valuable for music, where metadata is often incomplete, inconsistent, or missing, and where content is largely non-textual. However, unlike in most other domains, music items are multimodal, culturally situated, and frequently part of the long tail or newly released, limiting the effectiveness of general LLM approaches. As a result, LLM-based item modeling in music often requires additional grounding mechanisms, such as retrieval augmentation or multimodal alignment, to ensure accurate and up-to-date representations, especially for less popular or new content.

More broadly, general LLM-based item modeling methods do not transfer directly to music because they may fail to capture essential audio, contextual, or affective properties, and may propagate cultural biases, generic labels, or hallucinated attributes. When aggregating multiple sources, LLMs may also introduce inconsistencies or over-represent dominant cultural perspectives, which can negatively affect downstream recommendation and personalization. Music description must remain adaptable to different users and contexts, requiring controllable and personalized item representations rather than static tags or summaries.

Beyond subjective tags, even mid-level MIR descriptors such as key, tempo, or chord estimation remain unreliable, particularly for non-Western or underrepresented musical traditions, given the Western bias of most datasets [19, 57, 68, 116]. Thus, extracted features may be systematically inaccurate when applied to diverse repertoires. This issue is amplified in LLM-based pipelines, where generated descriptions often rely on upstream MIR outputs: if these outputs are flawed or culturally misaligned, the resulting captions may appear coherent yet misrepresent the music.

Concluding remarks on LLMs for Natural Language Music Recommendation. LLMs enable NL-based music recommendation by directly generating ranked item lists from user requests through prompting strategies such as zero-shot, few-shot in-context learning, retrieval-augmented generation (RAG), and reasoning scaffolds. These approaches allow flexible interaction, making them particularly promising for cold-start users, exploratory queries, and complex intents that combine factual constraints with subjective or affective descriptions. However, the effectiveness of these methods in

music depends strongly on how prompts are constructed and grounded. Zero-shot and few-shot approaches rely heavily on the model's prior knowledge or selected examples, which may over-represent popular, culturally dominant music or short-term user interests. Moreover, in music, other challenges are the ambiguity of entity names, the subjectivity and cultural dependence of descriptors such as mood or energy, and the sparsity and inconsistency of textual metadata. As a result, entity resolution and careful example selection become essential design components, and naïve transfer of general LLM prompting strategies may amplify noise, bias, or recency effects present in implicit listening logs.

Domain-specific adaptations of LLM-based NL recommendation often incorporate retrieval grounding and reasoning scaffolds, which improve transparency, controllability, and factual alignment with catalog content. However, applying RAG in music can be challenging as relevant information is fragmented across heterogeneous and often inconsistent sources (e.g., metadata, tags, reviews), and only a limited portion can be included in the prompt. Moreover, playlist generation introduces sequence-level constraints, where recommendation quality depends on transition coherence and overall flow. These domain-specific requirements mean that general LLM-based recommendation methods should not be directly applied without modification and careful consideration.

3 LLMs in Music Recommendation: Evaluation

In the previous section, we discussed the challenges associated with user modeling, item modeling, and NL-based recommendation in the music domain. We also summarized the key changes introduced by the adoption of LLMs in each of these areas. In what follows, we shift our focus to evaluating LLMs when used in any of the previously discussed scenarios. We begin by providing a comprehensive overview of how LLMs are evaluated for generative tasks, as designed and discussed within the NLP community (Section 3.1). We then extend the discussion to MRS-specific evaluation (Section 3.2), where we formalize universal success dimensions, discuss configuration-specific diagnostics, and illustrate the framework with working examples. Our goal is to provide an operational reporting framework and diagnostic toolbox for future empirical studies of LLM-based music recommendation.

3.1 NLP-Driven Evaluation

When using LLMs for user modeling, item modeling, or music recommendation, the task generally reduces to learning $P_\theta(y \mid x)$, a conditional probability distribution over output texts y (e.g., user preference profiles, playlist captions, music recommendations in natural language), given a textual or multimodal input x (e.g., track metadata, text-based listening history, tokenized audio). For simplicity, the task can also be expressed as an input-to-text mapping function as follows:

$$f : X \to Y, \quad x \in X,\ y \in Y,$$

where X is the set of possible inputs and Y is the set of possible output texts.

Evaluation strategies can focus either on comparing the predicted output y against a reference output y^* or human judgments (**reference-based evaluation**), or on comparing the predicted output y directly against the source input x (**reference-free evaluation**). Reference-based evaluation, which *relies on ground-truth data*, remains the most common approach. However, reference-free evaluation, often based on LLMs (e.g., LLMs-as-a-judge), has recently emerged as a promising, resource-efficient alternative, as it does not require extensive human annotation and is thus suitable for fast and large-scale evaluation experiments. Figure 4 summarizes the content that will be discussed in this section, presenting it as a decision process to guide the selection of the most suitable evaluation metrics based on the specific characteristics of each scenario.

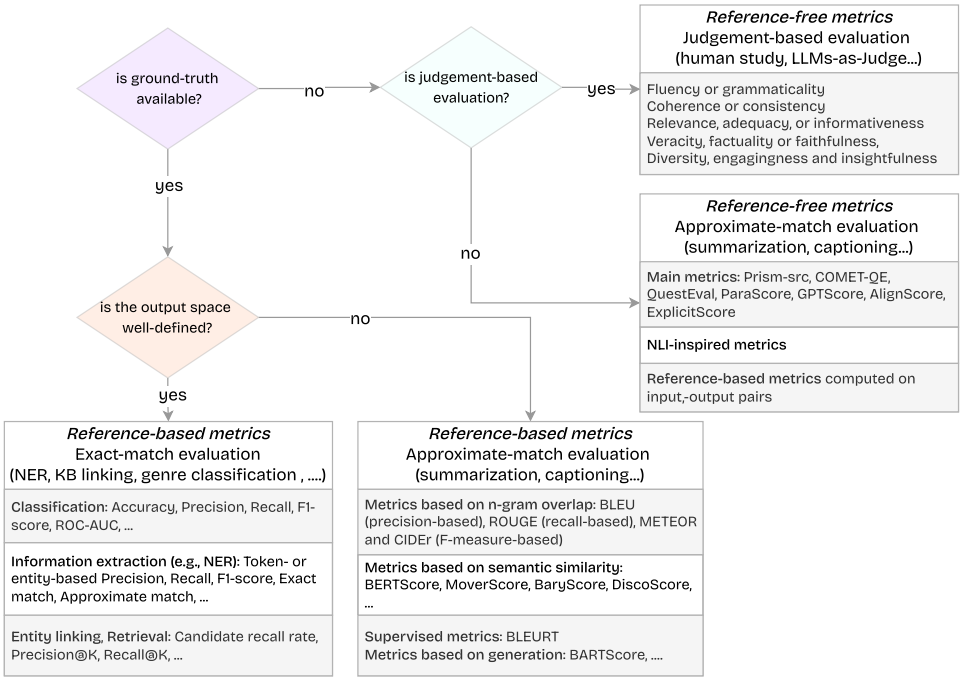


Fig. 4. NLP-driven evaluation framework: decision process and metrics overview.

In what follows, we set aside aspects related to content parsing. Modern LLMs can generate structured outputs directly (e.g., JSON files), which provides more suitable formatting of results for evaluation. Therefore, we do not consider evaluation metrics related to this parsing phase and focus only on the quality of the generated content itself.

3.1.1 Reference-based Evaluation. For the previously described roles of LLMs, the evaluation of success when references / ground-truth examples are available can be discussed as:

(1) **Exact Match Evaluation**, used when the output can be represented in a well-defined, discrete space with unambiguous ground truth (i.e., limited number of outcomes or outcomes expected to be limited to parts of input in case of information extraction). This includes tasks such as NER (i.e., matching predicted music entities to reference spans, either token-wise or entity-wise) [47, 71], emotion annotation (e.g., music classification into multiple discrete classes [202]), and knowledge base linking (e.g., checking whether a predicted item identifier corresponds exactly to the reference entity in the KB [142]). Formally, given a ground-truth output y^* and a model-generated output $y = f(x)$, the **evaluation metric** can be defined as a function:

$$M : Y^* \times Y \to \mathbb{R}, \quad Y^* \subseteq Y \quad \text{and} \quad |Y^*| \ll |Y|$$

which measures the degree of similarity or correspondence between the reference and predicted outputs:

$$M(y^*, y) \in \mathbb{R}.$$

As Y^* consists of a limited number of outcomes (e.g., a set of class labels, named entities, or knowledge base identifiers) in the exact match evaluation, M corresponds to exact-match success metrics such as **Accuracy**, **Precision**, **Recall**, or **F1-score**. Ranking metrics such as **Precision@k** or **Candidate Recall Rate** could also be used, for instance, when assessing the linking of an entity to a knowledge base [155].

(2) **Approximate Match Evaluation**, used when the output is free-form text or belongs to a large, open vocabulary where multiple correct answers exist. Examples include music text summarization [206], music captioning [60, 66], or user profile generation. Formally, when Y is the output space of open-form textual possibilities (i.e., any natural language text) with the ground-truth text $y^* \in Y$ for approximate match evaluation, the evaluation metric M can be defined as a quality estimation function between the reference and predicted texts:

$$M : Y \times Y \rightarrow \mathbb{R}, \quad M(y^*, y) = S(\phi(y^*), \phi(y)),$$

where $\phi(\cdot)$ denotes a transformation (e.g., embedding model, sparse feature extractor) that maps text into a continuous representation, and $S(\cdot, \cdot)$ is either a similarity measure based on distance alignment, frequency-representation of n-gram overlap, or the probability of y being generated from y^* when estimated with another pre-trained generative model. In this setting, higher $M(y^*, y)$ values indicate closer semantic alignment between the reference and generated texts. Typical choices of M for this type of evaluation include:

- metrics based on n-gram overlap (**BLEU** [133] — precision-based metric, **ROUGE** [102] — recall-based metric, **METEOR** [126], and **CIDEr** [190] — F-measure-based metrics),
- metrics based on semantic similarity (**BERTScore** [218], **MoverScore** [219], **BaryScore** [31], **DiscoScore** [220]),
- supervised metrics, i.e., regressive models on top of text encoder to predict human judgments (e.g., **BLEURT** [168]),
- metrics based on text generation (e.g., **BARTScore** [207])

An overview of these evaluation metrics, exemplified when used for item modeling, is presented in Table 4.

Although widely adopted in the community, the approximate match evaluation metrics have been increasingly scrutinized. Studies have shown that n-gram–based metrics often correlate poorly with human judgments [119, 140], while similarity-based metrics can be overly sensitive to adversarial perturbations or lexical variations [158]. To address these limitations, a new line of research [27, 82] has proposed designing evaluation metrics inspired by the **Natural Language Inference** (NLI) paradigm, where a model-generated output y is considered of high quality if it is semantically equivalent to a human reference y^* under the notion of *bi-implication* (i.e., logical equivalence). Furthermore, combining traditional metrics with NLI- or QA- (Question Answering) based approaches has been shown to yield more robust and reliable evaluation signals, improving correlation with human assessments [27, 70].

Overall, the evaluation of natural language generation with the **reference-based framework** remains an active and evolving area of research. Current findings underscore the need to develop automatic success metrics that align more closely with human judgments, particularly across complex and multi-dimensional aspects of text quality. Moreover, there has been relatively little investigation into how general-domain evaluation metrics perform in specialized contexts, such as domain- or task-specific applications in RSs, where the nature of generated text and evaluation criteria may differ substantially. In such settings, personalization, reflecting individual user preferences, tastes, and context, plays a crucial role in determining the perceived quality and relevance of generated content. This contrasts with most NLP text generation tasks, where personalization is typically less central to evaluation.

Table 4. Overview of evaluation metrics, exemplified through item modeling tasks with LLMs.

Task Type	Example Output	Common Metrics
Classification	"high valence" / "low valence"	Accuracy, Precision, Recall, F1, ROC-AUC
Extraction (NER)	track=Love, artist=The Beatles	Precision, Recall, F1-score (token- or entity-level), Exact Match, Approximate Match
Entity Linking	Knowledge-based identifier: `track:12345`	Accuracy, Precision, Recall, F1, Candidate Recall Rate, Precision@k
Summarization	"A warm, soulful album blending jazz and funk."	ROUGE, BLEU, METEOR, BERTScore, Human judgments (fluency, informativeness)
Captioning	"A mellow jazz ballad with soft piano and vocals."	BLEU, ROUGE, CIDEr, BERTScore, Human evaluation (naturalness, relevance)

3.1.2 Reference-free Evaluation. Because collecting human annotations for generative tasks is time-consuming and costly, a growing body of research has explored reference-free evaluation frameworks for tasks requiring an *approximate match evaluation*, when the model's output is evaluated with respect to the given input [27, 40]. In addition, these metrics are also considered valuable in online scenarios requiring real-time text quality estimation (e.g., captioning music playlists or answering questions about music in a conversation). Formally, when Y is the open-form output space (i.e., any natural language text, in the case of playlist captioning could be any music-related description), a reference-free evaluation metric M can be defined as a function of the input x (e.g., the metadata of the tracks in the playlist) and the model-generated output y (e.g., the caption), which can be viewed as an explicit or implicit conditional generation model [40, 183]:

$$M : X \times Y \to \mathbb{R}, \quad M(x, y) = P_\phi(y \mid x),$$

where ϕ denotes the parameters of the underlying evaluation model, and higher values of M indicate better-quality outputs. The underlying evaluation model P_ϕ is typically larger or more complex than the LLM used in recommendation, P_θ, but it also has its own limitations and, therefore, cannot be used to solve the task directly—for instance, it may not be specifically optimized for that recommendation task or could be prohibitively expensive to run. Reference-free metrics could employ evaluation models in a zero-shot setup (i.e., using pre-trained LLMs [54]), while others [151] undergo an optimization process based on some training data to estimate ϕ.

Examples of reference-free metrics, M, are **Prism-src** [183], **COMET-QE** [151], **QuestEval** [167], **ParaScore** [174], **GPTScore** [54], **AlignScore** [212], **ExplicitScore** [28], or **NLI-inspired** metrics [27]. Reference-based metrics for text generation tasks (e.g., **BARTScore** or **BERTScore**) can also be adapted for reference-free evaluation by computing them between the input x and the generated output y, provided x is itself textual [174].

Although reference-free metrics were initially found to underperform compared to reference-based ones, recent work has shown substantial improvements in their ability to approximate human judgments of text quality [53, 174]. However, their reliability remains inconsistent across different natural language generation tasks, setups, and domains, with correlations to human evaluation varying considerably depending on linguistic attributes, e.g., negation and sentiment polarity are poorly captured by embedding-based metrics in machine translation [53]. Also, lexical divergence in paraphrasing—especially when text is written by humans—is measured inadequately [174]. This variability highlights an important research challenge: designing generalizable, reference-free evaluation

methods that robustly capture human-perceived quality across diverse tasks, data distributions, and complex linguistic phenomena. Addressing this challenge requires a deeper understanding of how subjective dimensions of quality can be operationalized. Furthermore, extending such evaluation metrics to handle **multimodal inputs**, as often encountered in music-related applications (e.g., captioning or cross-modal summarization), remains an even more challenging yet crucial direction for future investigation.

3.1.3 Other Evaluation Approaches. As previously mentioned, **human evaluation** is widely regarded as the gold standard for assessing open-ended text generation tasks and automatic metrics seek to correlate with human judgments. Human evaluation is also useful when automatic evaluation through metrics is not able to capture more subjective or complex aspects of quality. Typical evaluation dimensions investigated in human studies of LLM-based MRSs include [160, 187]:

- **Fluency** or **grammaticality**, which measure how natural, readable, and error-free the output is;
- **Coherence** or **consistency**, which assesses if the text flows logically without contradictions;
- **Relevance**, **adequacy**, or **informativeness**, which reflects how well the generated text preserves important content from the source or fulfills the task;
- **Veracity**, **factuality**, or **faithfulness**, which analyzes if the output is accurate and grounded in the source material or external knowledge (this is also a measure of hallucination presence);
- **Diversity**, **engagingness**, and **insightfulness**, which captures originality and avoidance of repetition, as well as how interesting or likeable the output is (relevant in creative tasks such as music captioning).

Certain applications also employ task-specific criteria; for instance, **coverage and conciseness** in summarization [92].

Quality judgments are often collected through Likert scales (e.g., 1–5 or 1–7) or pairwise preference rankings, in which annotators directly compare two outputs. Significant efforts have also been made to assess the correlation between human judgments on these complex dimensions and existing automatic metrics [187], with mixed results. For example, **coherence** has been shown to correlate relatively well with n-gram-based metrics (e.g., BLEU, METEOR), whereas **insightfulness** often shows no correlation with established metrics [187].

However, as we mentioned before, given that the evaluation involving humans still requires significant efforts for collecting results on these subjective dimensions, more recent works employ **LLM-as-a-judge** [13, 29, 50, 193]. LLMs appear to be reliable evaluators in a large range of tasks. However, they also exhibit substantial inconsistency depending on the evaluated dimension (e.g., **engagingness** has low correlation with human judgment while **informativeness** correlates better), expertise level of human annotators (if compared with expert vs. non-expert human annotators), and if the assessed text is human- or LLM-generated [13].

Thus, LLMs should be rigorously validated against human judgments before being adopted as automatic evaluators. In the context of music and MRSs, this raises an open and important question: To what extent can LLM-based evaluators reliably and fairly assess music-related textual content, such as tags or generated descriptions, given the subjectivity and cultural variability inherent to musical preference and user perception? Addressing this question is required to ensure the reliability, bias control, and validity of LLM-driven evaluation in such personalized and domain-specific settings.

3.1.4 Discussion of Risks in LLM-generated Text Evaluation. Multiple studies from the NLP domain [27, 110, 127, 145, 145, 178, 181] have outlined other pitfalls in evaluating LLM-generated text, besides those mentioned before. We identify

risks stemming from the model itself and its intricacies, related to the characteristics of the generated output, and associated with the evaluation metrics.

Output-Related Risks. In generative tasks, where the output can take many possible forms, recent research in NLP has identified several evaluation-related risks. First, the length of generated text has been shown to strongly influence evaluation outcomes for both n-gram-based and semantic similarity metrics [145]. Because LLMs often generate substantially longer outputs than ground-truth references, automatic metrics may penalize otherwise relevant or high-quality generations due to length mismatches. Recent work shows that computing evaluation metrics on automatically summarized outputs, using simple extractive summarization methods, can significantly improve the alignment between automatic scores and human judgments. Second, Long et al. [110] show that all LLMs have significant biases across various output formats (e.g., multiple-choice question formats, text wrapping styles, or list structures) spanning a wide range of benchmarks. Several strategies to mitigate these biases have been proposed, including *(1) incorporating format templates or demonstrations in the prompt* and *(2) repeating format instructions multiple times within the prompt.*

Metrics-related Risks. Risks pertaining to the used metric include social and representational biases, such as those observed with BERTScore [181]. Reference-free metrics [27] also pose challenges: they can favor models similar to themselves, and thus, even penalize higher-quality human outputs, suggesting they are better suited as diagnostic tools instead of absolute measures of task performance. Moreover, when LLMs are employed as evaluators, they have been shown to exhibit position bias [193]. That is, the order in which candidate responses appear in the prompt can skew evaluation results, potentially making one model appear significantly superior to another. To mitigate this bias, Wang et al. [193] propose a calibration framework, which also shows closer alignment with human judgments.

LLMs and Decoding Strategies. Given an input, the same LLM can produce highly variable outputs depending on the decoding strategy used (e.g., greedy decoding—selecting at each step the token with the highest predicted probability [69], or nucleus sampling—sampling from the smallest set of tokens whose cumulative probability exceeds a threshold [77]), a phenomenon referred to as non-determinism in generation [178]. However, for practical and computational reasons, LLM evaluations are often conducted using a single input–output pair. A recent study [178] evaluating multiple LLMs across several benchmarks has shown a significant gap between greedy decoding and sampling-based decoding: greedy decoding often yields better results than the average performance obtained from multiple outputs generated through sampling strategies. Also, employing a best-of-N sampling strategy (selecting the best output among N generations for the same input) can enable even smaller LLMs to match or outperform larger models. These findings carry important implications for the evaluation of LLMs. First, the inherent non-determinism of generation introduces a source of variability that challenges the reliability and reproducibility of evaluation results: a single input–output pair cannot adequately represent model performance. Second, decoding choices can significantly alter performance scores, making cross-model comparisons difficult unless decoding settings are standardized or explicitly reported. Third, methods such as best-of-N sampling reveal that a model's potential quality can be much higher than what single-sample evaluations suggest, raising questions about what aspect of performance (average, best-case, or expected quality) should be measured.

3.2 MRS-Driven Evaluation

Music recommendation with LLMs inherits the evaluation challenges of both modern RSs and natural language generation while adding domain-specific complications such as extreme catalog scale and long-tailedness, frequent new

releases, multilingual metadata, name ambiguity (artists/tracks that overlap with common words), and playlist-first consumption where ordering and transitions matter (Table 3, Section 2.3). In this setting, classical ranking metrics alone are insufficient. Although non-existent or unresolved tracks should be penalized by the standard top-K metrics when they are evaluated directly against a fixed ground truth, aggregate relevance scores do not by themselves distinguish hallucinated items from irrelevant but valid items, nor do they capture failures that are removed by post-processing or hidden by averaging. An LLM-based system may thus, obtain acceptable relevance scores while still producing a catalog grounding failures, violating explicit constraints ("no live versions", "only 1960s", "avoid explicit lyrics"), or systematically under-serving languages, regions, or music communities that are central to a listener's identity.

3.2.1 Two Stakeholder Perspectives. A central issue in evaluating LLM-based MRSs is that the notion of a "good" recommendation depends on the stakeholder under consideration. Evaluation should thus make explicit which utility the system is intended to optimize. Music streaming is a prototypical *multi-stakeholder* setting: listeners, creators (artists), publishers (record labels), and platform providers often pursue different, and sometimes conflicting, objectives. These tensions are particularly visible in sequential listening and discovery-oriented scenarios [2, 7, 42]. The most commonly targeted stakeholders in research are listeners and creators. Thus, we detail their usual objectives next.

From a **listener** perspective, evaluation metrics are proxies for distinct facets of user value. First, **short-term satisfaction** asks if the system follows the user request *in the current interaction*: it should return *valid catalog entities* and respect explicit constraints (e.g., "no live versions", "only 1960s", "avoid explicit lyrics"), since hallucinated or misattributed recommendations are failures even when the response is fluent [88, 135]. Second, **sustainable long-term value** (and discovery) ask if the system supports music discovery without extra focusing on overly popular content, reflecting the well-known exploration–exploitation tension between short-term satisfaction and long-term coverage, learning, novelty, and diversity [89, 189]. Third, **personalization value**, **emotional well-being**, and **psychological effect** ask if recommendations feel genuinely tailored to the listener, add value beyond a one-size-fits-all or popularity-driven experience, and respond to affective needs and listening intents such as mood regulation, stress reduction, emotional expression, and social connectedness, thereby supporting overall well-being [161, 215]. Fourth, **agency**, **identity fit**, and **inclusion** ask if the system reflects user identity (e.g., language, scene, region, or cultural background) and provides sufficient controllability so that users can scrutinize and steer recommendations as tastes evolve [76, 144, 148]. This point is particularly important for listeners with non-mainstream or minority preferences, who may otherwise be systematically underserved by popularity-biased systems [91, 95]. Finally, **trust** and **acceptance** are shaped not only by relevance but also by explanation quality: in music recommenders, explanations can improve trust, forgiveness, and user tolerance for unexpected suggestions, especially when recommendations are surprising but potentially valuable [4, 184].

From a **creator** and **rightsholder** perspective, evaluation must also capture **attribution integrity** and **exposure balance**. Attribution integrity refers to if recommended works are correctly linked to the corresponding artists, tracks, albums, and rights holders. Exposure balance refers to if recommendation attention is not concentrated only on already popular artists, but also reaches long-tail, emerging, local, and culturally or linguistically diverse parts of the catalog. Recommendation is thus a form of attention allocation: as exposure is position-dependent, ranked recommendations can create downstream economic and cultural impacts for creators and provider groups. Fairness-aware evaluation should thus consider disparities across artist, label, provider, cultural, and linguistic segments [42, 176].

From a **platform-provider** perspective, evaluation should also report **operational** and **policy robustness**. Platforms operate under non-functional constraints, such as latency, cost stability, scalability, and catalog freshness, as well

as governance and regulatory constraints, such as content moderation, safety, and compliance. These system-level concerns should be reported alongside user-facing effectiveness instead of being conflated with relevance or satisfaction metrics [209, 221].

Because these objectives interact, we recommend *multi-objective reporting* rather than collapsing evaluation into a single scalar score. For example, increasing long-tail exposure may improve discovery and creator-side exposure balance, but it may also reduce short-term hit-rate for some users. Similarly, stricter policy filtering may improve safety while reducing catalog coverage. Multi-objective reporting makes such trade-offs explicit and is better aligned with the inherently multi-stakeholder nature of music recommendation [2, 7, 36]. More broadly, evaluation should ideally be informed by feedback from diverse affected parties instead of a single optimization target alone.

3.2.2 Universal Evaluation Dimensions and Metrics. We define success in LLM-based music recommendation as *faithfully satisfying a user request with real catalog items while supporting discovery, personalization, controllability, and cultural reach*. We introduce an MRS-oriented evaluation framework centered on multiple dimensions relevant across all LLM configurations, zero-shot, profile-conditioned, ICL-based, retrieval-augmented, or reasoning-based (Section 2.3). These dimensions answer six different questions:

G1. Query adherence and groundedness. Does the system return real catalog entities, and does it respect the user's explicit constraints? This is the first and most minimal requirement for LLM-based recommendation, and it captures the **short-term satisfaction** aspect.

G2. Discovery quality. When the user intent is exploratory, does the system promote discovery—for example, by increasing novelty and diversity and by allocating exposure beyond the head of the popularity distribution? This is a canonical "beyond-accuracy" objective, closely related to coverage and **long-term value**.

G3. Personalization uplift. Does conditioning on the user profile or listening history improve recommendation utility compared with an otherwise identical non-personalized baseline, such as a generic, popularity-based, or request-only recommender? Here, "uplift" refers to the marginal gain attributable to personalization, for example a difference in nDCG@K, QueryAdherence@K, user-relative novelty, or human preference scores between profile-conditioned and profile-ablated recommendations.

G4. Profile fidelity and controllability. Do recommendations match the user's stated preferences, and do they change *predictably* when the user edits or refines their preference representation? This operationalises the **user-agency dimension**.

G5. Cultural and linguistic coverage. Does the system represent the languages, regions, and music communities relevant to the request fairly and appropriately? Here, music communities refer to socially or geographically situated scenes or subcultures, such as local punk scenes, Detroit techno, UK grime, or regional metal scenes. This dimension captures if the group composition of the recommended list, for example its distribution over languages, regions, or music communities, matches the target composition implied by the user request or by a stated evaluation policy.

G6. Classical relevance. How relevant are the recommended items under established top-K effectiveness metrics? This dimension may become especially relevant when historical user interactions are converted into natural-language profiles to drive recommendations. It serves as the compatibility layer with standard recommender evaluation, but it should be interpreted jointly with G1–G5 in LLM-based MRSs.

Table 5. Stakeholder-perspective interpretation of the universal success dimensions. The mapping is intentionally many-to-many: some goals are primarily captured by one dimension, whereas others require a joint reading of several dimensions. Operational and policy robustness remain important platform concerns, but they should be reported alongside rather than inside G1–G6.

Stakeholder goal	Primary dimension(s)	Interpretation in the evaluation framework
***Listener*: request satisfaction and trust**	**G1**, **G6**	G1 checks if the system returns real catalog items, follows explicit constraints, and explains the list faithfully, while G6 adds the conventional top-K relevance layer. Together they capture the minimal conditions for *short-term satisfaction* and acceptance.
***Listener*: discovery and sustained value**	**G2**, **G3**	G2 measures long-tail exposure, lower popularity concentration, diversity, and catalog breadth. G3 complements this through user-relative novelty and the incremental benefit of profile conditioning beyond generic behavior.
***Listener*: personalization, agency, and affective fit**	**G3**, **G4**	G3 asks if profile conditioning adds measurable utility, whereas G4 asks if recommendations remain faithful to the user representation and respond predictably to edits thereof. For mood- or activity-sensitive requests, this is also where affective fit becomes visible.
***Listener*: cultural and linguistic fit**	**G5**	G5 evaluates if the distribution of the recommended list across languages, regions, and music communities aligns with the target implied by the user request or a stated evaluation policy. Here, music communities include scenes or subcultures such as local punk scenes, Detroit techno, UK grime, or regional metal scenes. The metric also assesses if failures are evenly distributed across such groups.
***Creator/platform*: attribution integrity and faithful explanation**	**G1**	Entity resolvability verifies that recommended items correspond to real catalog entities. Explanation–list consistency verifies that the rationale is faithful to the returned list and does not introduce unsupported attributes, artists, tracks, or reasons.
***Creator/platform*: exposure balance and ecosystem health**	**G2**, **G5**	G2 captures long-tail exposure, popularity concentration, and catalog breadth, while G5 captures representation balance across languages, regions, or scenes. Together, they address exposure allocation beyond pure relevance.
***Creator/platform*: operational and policy robustness**	*Reported alongside G1–G6*	Latency, cost stability, safety, moderation, and regulatory compliance are important system-level concerns, but they are not themselves universal success dimensions in the present framework.

The introduced dimensions are complementary. Running Example 1 (see Section 3.2.3) shows that a system may be well grounded under G1 while still failing to satisfy the actual request. Running Example 2 (see Section 3.2.3) shows that a system may achieve positive personalization gain under G3 while still failing discovery under G2. Accordingly, the role of the universal metrics is not to collapse evaluation into a single scalar score, but to expose distinct aspects of recommendation quality that matter in music recommendation. Table 5 makes this explicit by showing how the listener- and creator/platform-side goals discussed above map onto the operational success dimensions G1–G6, while also clarifying that some system-level concerns, such as operational and policy robustness, should be reported alongside instead of being conflated with success metrics.

Next, we operationalize the evaluation framework by making the dimensions explicit, defining the evaluation unit, and distinguishing strict feasibility constraints from graded satisfaction of fuzzy music descriptors. Several metrics are query-conditioned; their desired direction and interpretation should thus be assessed relative to the user intent and recommendation setting.

G1. Query adherence and groundedness. Let u denote a user, q a natural language request at a given conversational turn, I the music catalog, and $R_K(q,u) = (i_1, \ldots, i_K)$ the top-K items. Unless otherwise stated, metrics can be computed per request and macro-averaged over an evaluation set Q. Let $\mathrm{res}(i) \in \{0,1\}$ indicate if item i can be resolved to a

canonical catalog entity (e.g., a MusicBrainz identifier). We define:

$$\text{EntityGroundedness@K}(q) = \frac{1}{K}\sum_{j=1}^{K} \text{res}(i_j).$$

which measures the proportion of top-K recommendations that can be successfully matched to entities in the target catalog, thereby capturing a fundamental validity requirement of free-text generative RSs [37, 74]. Higher values indicate that the system produces executable and catalog-grounded recommendations rather than fabricated or unusable outputs. From the **listener side**, high values support *short-term satisfaction* and *trust*, since recommended items can actually be played or verified. From the **creator/platform side**, they help preserve *attribution integrity* and *reliable catalog coverage*. Conversely, low values point to hallucination-related failures (see below) that weaken usability, trust, and eventually long-term engagement [88].

We define the hallucination rate as the complement of entity groundedness:

$$\text{HallucinationRate@K}(q) = 1 - \text{EntityGroundedness@K}(q).$$

This metric measures the proportion of top-K recommendations that cannot be resolved to entities in the target catalog, thereby capturing unsupported or non-existent recommendations [37, 88]. Lower values are desirable, since hallucinated items are one of the most visible failure modes of LLM-based recommendation.

We next separate strict query feasibility from graded query alignment. Let $C(q) = \{c_1, \ldots, c_m\}$ denote the set of requirements to be evaluated extracted from query q. These requirements may include strict constraints, such as "no live versions" or "released in the 1990s," as well as graded descriptors, such as "not too mainstream" or "similar to this, but calmer." Let $C_{\text{hard}}(q) \subseteq C(q)$ denote the subset of non-negotiable constraints. For each requirement c_ℓ, let $s_\ell(i) \in [0,1]$ denote the degree to which item i satisfies it, where $s_\ell(i) = 1$ indicates full satisfaction. Following constraint-based recommendation and recent instruction-following evaluation, we first define the hard satisfaction predicate as [1, 51]:

$$\text{Sat}_{\text{hard}}(i, q) = \prod_{c_\ell \in C_{\text{hard}}(q)} \mathbf{1}[s_\ell(i) = 1],$$

with the convention that the empty product equals one when no strict constraint is specified. Strict query adherence is then defined as:

$$\text{QueryAdherence@K}(q) = \frac{1}{K}\sum_{j=1}^{K} \text{res}(i_j)\,\text{Sat}_{\text{hard}}(i_j, q).$$

This metric measures the proportion of top-K recommendations that are both executable catalog items and satisfy all non-negotiable query constraints. It is thus a strict feasibility measure: for a request such as "Brazilian metal, no live versions," an item contributes positively only if it resolves to a valid catalog entity, is compatible with the requested style or origin, and is not a live recording. High values indicate strong request fidelity under explicit constraints, whereas low values indicate failures of grounding, constraint following, or both.

For subjective, fuzzy, or partially satisfiable descriptors, a strict binary predicate is often too severe. We thus define a weighted soft aggregation operator:

$$A_\alpha\big(s_1(i), \ldots, s_m(i)\big) = \sum_{\ell=1}^{m} \alpha_\ell s_\ell(i), \qquad \alpha_\ell \geq 0, \qquad \sum_{\ell=1}^{m} \alpha_\ell = 1.$$

The corresponding soft query alignment score is:

$$\text{QueryAlignment@K}(q) = \frac{1}{K}\sum_{j=1}^{K} \text{res}(i_j)\, A_\alpha\big(s_1(i_j), \ldots, s_m(i_j)\big).$$

This metric measures the average degree to which the returned list matches the full set of query requirements, while allowing partial satisfaction. In contrast to QueryAdherence@K, it remains informative when the user request contains fuzzy musical descriptors, subjective affective terms, or trade-offs among multiple facets [1, 51]. For instance, a track that matches the requested genre and decade but only partially matches the desired mood can still receive partial credit. The factor $\text{res}(i_j)$ prevents unresolved items from receiving alignment credit; a semantic-only variant can omit this factor, but should then be reported separately from catalog-grounded alignment.

To diagnose constraint-following independently of entity resolution, we also define the hard constraint hit rate:

$$\text{ConstraintHitRate@K}(q) = \frac{1}{K}\sum_{j=1}^{K} \text{Sat}_{\text{hard}}(i_j, q).$$

This metric isolates if the recommended items satisfy the strict semantic requirements of the query, regardless of whether they can be resolved in the target catalog. Read together, QueryAdherence@K, QueryAlignment@K, and ConstraintHitRate@K distinguish three cases: fully valid recommendations, near misses that partially match the intent, and failures caused specifically by catalog grounding.

Finally, let T_q denote the explanation text associated with query q, and let $\mathcal{M}(T_q) \subseteq \mathcal{I}$ be the set of normalized catalog entities mentioned in that explanation. We define explanation groundedness as:

$$\text{ExplanationGroundedness}(q) = \frac{\big|\mathcal{M}(T_q) \cap \{i_1, \ldots, i_K\}\big|}{\max\{1, |\mathcal{M}(T_q)|\}}.$$

This metric measures if the explanation refers to items that actually appear in the returned recommendation list, thus capturing a basic form of explanation faithfulness [86, 185]. Higher values indicate that the rationale is grounded in the recommended list rather than introducing unsupported artists, tracks, or attributes. Low values signal a mismatch between the recommendation and its explanation, even when the list itself contains valid items.

G2. Discovery quality. Let $\text{pop}(i) \in [0, 1]$ denote a normalized popularity score, and let $\text{LT}_\tau(i) = \mathbf{1}[\text{pop}(i) \le \tau]$ indicate if item i belongs to the long tail under threshold τ, we define

$$\text{LongTailShare@K}(q) = \frac{1}{K}\sum_{j=1}^{K} \text{LT}_\tau(i_j).$$

which measures the proportion of top-K recommendations drawn from the long tail of the catalog, following standard beyond-accuracy evaluations of popularity bias and exposure concentration [3, 136]. Within the discovery quality dimension, higher values indicate a stronger tendency to expose users to less popular and potentially under-discovered content. From the **listener side**, high values suggest greater opportunity for exploration beyond overexposed content, while from the **creator/platform side**, they indicate broader long-tail exposure and a fairer distribution of attention across the catalog. Importantly, the interpretation of this metric remains query-dependent: high long-tail exposure is particularly desirable in exploratory or discovery-oriented scenarios, but may be less appropriate for requests explicitly targeting canonical, trending, or mainstream content.

We then define the average recommendation popularity as:

$$\text{ARP@K}(q) = \frac{1}{K}\sum_{j=1}^{K}\text{pop}(i_j),$$

This metric measures the average popularity level of the recommended items and therefore provides a direct indicator of popularity concentration and head-catalog exposure [3]. Within the discovery quality dimension, lower values generally indicate a greater tendency to recommend less dominant and potentially less overexposed content. For example, a recommendation list composed primarily of globally popular or canonical tracks will yield a high ARP@K. From the **listener side**, lower values tend to support exploration and reduce repetitive mainstream exposure. From the **creator/platform side**, they indicate a less concentrated distribution of exposure across the catalog. Conversely, high ARP, especially when combined with low long-tail share, usually signals weaker discovery quality, reduced catalog breadth, and more concentrated creator-side exposure.

Let $d(i_a, i_b) \in [0, 1]$ denote a pairwise dissimilarity measure between items i_a and i_b. We define intra-list diversity as the average pairwise dissimilarity among the recommended items, according to some dissimilarity measure computed on feature representations of i_a and i_b:

$$\text{ILD@K}(q) = \frac{2}{K(K-1)}\sum_{1\le a<b\le K} d(i_a, i_b).$$

This metric is the standard list-level diversity measure used in RSs evaluation [189, 226]. Higher values indicate that the recommendation list covers a broader range of items. For example, a recommendation list spanning different artists, moods, genres, or audio characteristics will typically exhibit higher ILD than a list composed of closely related tracks. From the **listener side**, high ILD reduces redundancy and helps keep exploratory sessions engaging. From the **creator/platform side**, it broadens exposure across different parts of the catalog. At the same time, ILD should still be interpreted jointly with relevance, so that diversity is not achieved simply by drifting away from the user's request.

For system-level metrics, let

$$\mathcal{Q} = \{e_n\}_{n=1}^{N}, \qquad e_n = (q_n, u_n, \mathcal{Y}_n^\star, C(q_n), \pi(q_n)),$$

denote an evaluation set of recommendation cases. Here, q_n is the natural-language request, u_n is the associated user when personalization is evaluated, $\mathcal{Y}_n^\star$ denotes optional ground-truth relevant items or relevance labels, $C(q_n)$ is the set of constraints extracted from the request, and $\pi(q_n)$ is an optional target group distribution used for cultural or linguistic coverage metrics. For reference-free metrics, $\mathcal{Y}_n^\star$ may be empty.

For an evaluation set $\mathcal{Q}$, we define system-level catalog coverage as

$$\text{CatalogCoverage@}K(\mathcal{Q}) = \frac{\left|\bigcup_{e_n\in\mathcal{Q}} R_K(q_n, u_n)\right|}{|I|}.$$

which measures the fraction of the available catalog that is surfaced by the recommender across a set of requests, following standard beyond-accuracy evaluation in RSs [75, 189]. Higher values indicate broader catalog utilization and reduce the risk that the system repeatedly concentrates recommendations on a narrow subset of highly exposed items. For example, if a recommender repeatedly cycles through the same small set of artists, catalog coverage will remain low even when individual recommendation lists appear relevant. From the **listener side**, higher coverage supports *long-term satisfaction* by reducing stagnation and enabling broader exploration. From the **creator/platform side**, it

indicates healthier exposure across the inventory and less concentration of attention. Conversely, low coverage points to weaker discovery and more limited catalog use over time.

G3. Incremental value of personalization. Let $U(\cdot)$ denote a utility functional (e.g., nDCG@K, QueryAdherence@K, or a human preference score). $R(q, u)$ denotes the recommendation list generated with user profile conditioning, while $R(q, \emptyset)$ denotes the corresponding non-personalized recommendation list generated without user-profile information. For user u, we define the profile uplift as:

$$\Delta_{\text{profile}}(u) = \frac{1}{|Q_u|} \sum_{q \in Q_u} \left(U(R(q, u)) - U(R(q, \emptyset)) \right),$$

where Q_u is the evaluation set associated with user u. This metric measures the incremental value contributed by profile conditioning rather than recommendation quality in isolation, and thus serves as a central metric for the personalization value dimension [75]. From the **listener side**, a positive uplift means that the system contributes genuine personalization value. From the **creator/platform side**, it suggests better audience matching and less wasted exposure. Conversely, low or negative values indicate weaker personalization value, lower perceived care, and potentially reduced trust and long-term engagement. We further define the inter-user overlap as:

$$\text{InterUserOverlap@K} = \mathbb{E}_{u \neq u'} \left[\frac{|R_K(q, u) \cap R_K(q, u')|}{|R_K(q, u) \cup R_K(q, u')|} \right],$$

where the recommendation lists are interpreted as sets when computing the Jaccard overlap. This metric measures the average overlap between recommendation lists across different users, thereby diagnosing personalization collapse and insufficient user differentiation [84, 96]. Lower values generally indicate stronger personalization, since they suggest that the system adapts recommendations to individual user characteristics instead of repeatedly returning the same items to everyone. For example, if two users receive identical top-10 recommendation sets, then their pairwise overlap is equal to 1. From the **listener side**, lower overlap suggests that recommendations are less generic and more identity-specific. From the **creator/platform side**, it supports more precise audience allocation and broader differentiation of exposure. Conversely, persistently high overlap indicates weaker personalization value and poorer identity fit, even when the returned lists appear superficially plausible.

Next, let $H_u \subseteq I$ denote the prior exposure history of user u (preferably exposure rather than consumption alone in implicit-feedback settings). We define user-relative novelty as:

$$\text{UserNovelty@K}(q, u) = \frac{1}{K} \sum_{j=1}^{K} \mathbf{1}[i_j \notin H_u].$$

which measures the proportion of recommended items that are new relative to the user's prior exposure history, corresponding to the standard user-relative notion of novelty in RS evaluation [75, 189]. Higher values indicate that the recommender surfaces content that extends beyond the user's previously encountered items. From the **listener side**, high novelty supports discovery and sustained engagement. From the **creator/platform side**, it broadens exposure to content that is new to the user instead of repeatedly reallocating attention to already familiar items. Conversely, low novelty gain suggests that even if profile conditioning improves other metrics, the recommender may still fail to expand the user's effective music space and thus contribute only weakly to *long-term engagement.*

G4. Profile fidelity and controllability. For profile-based metrics, let H_u denote the interaction history of user u, and let

$$R_K(q, u) = (i_1, \ldots, i_K)$$

denote the top-K recommendation list returned for request q and user u. Let $\Phi(H_u)$ and $\Phi(R_K(q,u))$ denote smoothed facet distributions induced by the user's interaction history and by the recommendation list, respectively, over a fixed facet space $\mathcal{F}$, such as genre, decade, language, or mood. Thus, $\Phi(\cdot)$ maps a history or recommendation list to a normalized distribution over facets. Let $\mathrm{JSD}_2(\cdot,\cdot) \in [0,1]$ denote the base-2 Jensen–Shannon divergence. We define profile alignment as

$$\mathrm{ProfileAlignment@}K(q,u) = 1 - \mathrm{JSD}_2\left(\Phi(H_u), \Phi(R_K(q,u))\right).$$

This metric operationalizes profile fidelity as calibration between the user's facet distribution of historic interactions and the facet distribution induced by their recommendation list, following prior work on calibrated recommendation [103, 179]. Higher values indicate that the recommended items more closely reflect the user's stated or inferred preference profile across the selected facet space. For example, if the recommendation list closely matches the user's genre, language, and decade distributions, then ProfileAlignment@K approaches 1. From the **listener side**, high values support personalization, identity fit, and trust in the recommender. From the **creator/platform side**, they improve audience matching and reduce irrelevant exposure. Conversely, low values suggest weaker personalization value and poorer profile fidelity, and in requests involving mood or activity, they may also indicate reduced affective appropriateness.

Let e denote an explicit user edit (e.g., "same mood, but instrumental"), and let $q \oplus e$ denote the modified recommendation request obtained after applying the edit. We define edit success as the proportion of returned items that satisfy the updated constraint set:

$$\mathrm{EditSuccess@K}(e) = \mathrm{ConstraintHitRate@K}(q \oplus e).$$

This metric measures if the recommender successfully implements an explicit user intervention [26, 114]. High values indicate stronger controllability and more reliable interactive personalization, since the recommendation list adapts predictably in response to user refinements or corrections. For example, if an edit requests "less explicit lyrics" and 6 of the next 10 recommended items satisfy the updated hard constraints, then EditSuccess@10 = 0.6. From the **creator/platform side**, it means fewer wasted turns and resources, and more efficient user guidance. Conversely, low edit success implies that the user must invest more interaction effort before the system behaves as intended, which weakens both *short-term satisfaction* and the perceived value of conversational control.

Let $q_\tau^{(e)}$ denote the τ-th recommendation turn following an explicit user edit e. For a declared profile-fidelity threshold $\gamma \in (0,1]$, we define the adaptation latency as:

$$\mathrm{AdaptationLatency}(e;\gamma) = \min\left\{\tau \geq 1 : \mathrm{ProfileAlignment@K}(q_\tau^{(e)}, u) \geq \gamma\right\},$$

with value $+\infty$ (or a right-censored horizon) if the threshold is never reached. This metric measures how quickly the recommender adapts to the updated preference state induced by the user edit, and therefore serves as the conversational-efficiency counterpart of profile fidelity [61, 114]. Lower values are preferable, since they indicate that the system aligns with the edited user state in fewer interaction turns. For example, if the recommender requires three recommendation turns before the recommendation profile reaches the desired alignment threshold, then AdaptationLatency = 3. From the **listener side**, low latency reduces friction and supports more fluid interaction. From the **creator/platform side**, it means fewer wasted turns and more efficient user guidance.

G5. Cultural and linguistic coverage. For cultural and linguistic evaluation, we consider groupings that are relevant to the request or evaluation policy, such as language, region, country, or music community. Here, *music community* refers to a socially, geographically, or stylistically situated group of musical practice, such as Detroit techno, UK grime, Brazilian forró, or a local punk community. Let $\gamma(i) \in \mathcal{G}$ assign item i to one such group under the chosen grouping

Table 6. Summary of universal success dimensions and setting-specific diagnostics for LLM-driven music recommendation. The first block lists the universal metric families. The second block lists diagnostic bundles for ICL, RAG, and reasoning-based settings; these do not replace the universal layer, but explain configuration-specific variation around it.

Dimension / setting	Representative metrics	What it captures
Universal success dimensions (Section 3.2.2)		
G1. Query adherence and groundedness	EntityGroundedness@K, QueryAdherence@K, QueryAlignment@K, ConstraintHitRate@K, ExplanationGroundedness (paired complement: HallucinationRate@K)	Checks if returned items are real, query-faithful, and explained consistently. Higher values are preferred for the success metrics, whereas HallucinationRate@K should decrease.
G2. Discovery quality	LongTailShare@K, ARP@K, ILD@K, CatalogCoverage@K	Captures exploration through long-tail exposure, lower popularity concentration, within-list diversity, and broader catalog use. The desired direction is query-conditioned, but in exploratory settings LongTailShare, ILD, and coverage should increase, whereas ARP should decrease.
G3. Incremental value of personalization	Δ_{profile}, InterUserOverlap@K, UserNovelty@K	Measures if profile conditioning adds utility beyond generic behavior, differentiates users, and surfaces items that are new relative to prior exposure. Higher Δ_{profile} and UserNovelty@K are desirable, whereas lower InterUserOverlap@K indicates stronger personalization.
G4. Profile fidelity and controllability	ProfileAlignment@K, EditSuccess@K, AdaptationLatency	Measures fidelity to the user representation and responsiveness to explicit edits. Higher alignment and edit success are desirable, while lower latency indicates better controllability.
G5. Cultural and linguistic coverage	Share@K, TargetDisparity@K, WorstGroupGap@K	Evaluates cultural fit at the list level and robustness across groups. Share@K should track the intended target composition, while target disparity and worst group gap should decrease.
G6. Classical relevance	Precision@K, Recall@K, DCG@K, nDCG@K	Provides the classical relevance layer and remains useful as a compatibility baseline, but should be interpreted jointly with G1–G5 rather than on its own.
Setting-specific diagnostics (Section 3.2.4)		
S-ICL (few-shot prompting)	Shot calibration curves, PromptSensiScore, Rank-Biased Overlap, Preference injection tests, Instruction dominance over noisy/conflicting examples	Diagnoses if few-shot examples improve performance in a stable and controllable way, or instead introduce prompt brittleness and instruction drift.
S-RAG (retrieval-augmented)	EvidenceGrounding@k, Doc-swap test: Δ_{doc}, Recall@M, TTFS (Time-to-First-Surface), Day-0 Lift for new releases, Token cost	Diagnoses if retrieved evidence is actually used, if retrieval supports freshness, and what efficiency costs are incurred. These diagnostics should be read together with the universal grounding metrics in G1.
S-CoT / ToT (reasoning-based)	Reasoning faithfulness, Utility gain over direct decoding (ΔU_{CoT}), Flow coherence, Self-consistency uplift	Diagnoses if explicit reasoning improves recommendation quality, remains faithful to catalog evidence, and supports coherent sequence construction when ordering matters.

scheme. For a group $g \in \mathcal{G}$, we define its observed share in the recommendation list as:

$$\text{Share@K}(g; q) = \frac{1}{K} \sum_{j=1}^{K} \mathbf{1}\left[\gamma(i_j) = g\right].$$

This metric measures the proportion of the top-K recommendations that belong to group g [176, 211]. It is a descriptive composition measure rather than a success criterion by itself: its interpretation depends on the target implied by the user request or by the evaluation policy. For example, for the request "Brazilian forró with accordion solos," a high share of Brazilian or Portuguese-language items would be expected. By contrast, for a request asking for a balanced mix across Latin American regions, the observed shares should be compared against the desired target distribution.

Next, let $\pi_g(q)$ denote the target group distribution implied by a user request or by a declared evaluation policy such as a quota. We define disparity from the target composition as:

$$\text{TargetDisparity@K}(q) = \sum_{g \in \mathcal{G}} \left|\text{Share@K}(g; q) - \pi_g(q)\right|.$$

This metric measures the discrepancy between the observed group composition of the recommendation list and the intended target distribution, following fairness-of-exposure and equity-of-attention formulations [15, 176]. Lower values indicate that the returned recommendations more closely match the desired cultural or linguistic balance. For example, if a bilingual request asks for a balanced Spanish–Portuguese mix but the returned list is almost entirely Spanish, then the target disparity score will be high. From the **listener side**, low disparity means that the recommendation

Table 7. Schema of TalkPlayData–2 dataset.

Variable	Type	Description
`session_id`	string	A string identifier for the session (14–18 characters), typically combining the user ID and the date in the form `<userID>__<date>` (e.g., `53726__2011-12-26`).
`user_id`	string	A string identifier (2–6 digits) for the synthetic user participating in the session. This value can be used to relate multiple sessions involving the same user.
`session_date`	string	A date string (`YYYY-MM-DD`) indicating when the conversation took place. Dates in the dataset span roughly from 2006-05-28 to 2018-12-31.
`user_profile`	dict	A dictionary describing demographic and preference attributes for the user. Example keys include `age`, `age_group`, `country_code`, `country_name`, `gender`, `preferred_language`, `preferred_musical_culture`, and `user_split` (e.g., "train_warm").
`conversation_goal`	dict	A structured description of the listener's objective. It contains a category code (A–K), a free-text description of the goal (`listener_goal`), and a specificity label (LL, HL, LH or HH). The categories correspond to discovery mechanisms such as audio-based, lyrical, mood-based, etc., and the specificity dimension indicates how precise the request and the target recommendation are.
`conversations`	list	A list of dialog turns. Each element is a dictionary containing the `content` (either natural-language text or a track ID), the `role` (e.g., "user", "music" or "assistant"), an optional `thought` field representing the agent's reasoning, and the `turn_number`. This sequence captures the interaction history and the recommended tracks for the session.
`goal_progress_assessments`	list	A list that assesses how each turn moves toward the conversation goal. Each entry includes a `goal_progress_assessment` label (such as "MOVES_TOWARD_GOAL" or "DOES_NOT_MOVE_TOWARD_GOAL") and the corresponding `turn_number`.

list better matches the cultural or linguistic composition implied by the request. From the **creator/platform side**, it indicates less systematic underexposure of relevant groups. Conversely, high disparity suggests that the system may still appear relevant in a narrow sense while failing to provide cultural fit, identity recognition, or equitable exposure for underrepresented languages, regions, or scenes.

Let $Q_g \subseteq Q$ denote the evaluation slice associated with group $g \in \mathcal{G}$. We define the slice-level failure rate as:

$$F_g = 1 - \frac{1}{|Q_g|} \sum_{q \in Q_g} \text{ConstraintHitRate@K}(q).$$

We then define the worst-group gap as:

$$\text{WorstGroupGap@K} = \max_{g \in \mathcal{G}} F_g - \min_{g \in \mathcal{G}} F_g.$$

This metric measures how unevenly recommendation failures are distributed across groups, and is thus closely related to worst-group robustness under distribution or group shift [157, 176]. Lower values indicate that no particular user group experiences systematically worse recommendation quality relative to others. For example, if one language slice exhibits a failure rate of 0.45 while another exhibits a failure rate of 0.10, then the worst-group gap is 0.35. From the **listener side**, low values suggest that minority or non-dominant users are not substantially underserved. From the **creator/platform side**, they indicate fairer cross-group performance and a more balanced distribution of exposure. Conversely, a large gap means that average performance may still look acceptable while at least one group experiences persistently weaker adherence, thereby weakening inclusion, identity fit, and long-term trust in the platform.

Table 8. Turn–by–turn evolution of the running example. The table distinguishes contextual turns from scored turns and makes explicit which request is being evaluated at each stage.

Turn	User request summary	System response	Role in evaluation
1	Cover–art search only: the user remembers a striking, unsettling rock album cover, but does not yet specify genre, era, or exclusion constraints.	Talking Heads track (`1i6N76fftMZhijOzFQ5ZtL`); assistant explicitly recommends *Psycho Killer*.	Context only. This turn is not scored under the present G1 setup because the request is still too underspecified for strict constraint extraction.
2	The user clarifies that the goal is to identify an album from its unusual cover art, not to continue receiving Talking Heads recommendations.	Talking Heads track (`38Ngied9rBORlAbLYNCl4k`); assistant recommends *Once in a Lifetime*.	Context only. This turn still lacks a stable, checkable metadata constraint set, so it is not included in the quantitative G1 average.
3	The user adds explicit metadata constraints: *alternative or industrial rock* and *late 1990s*; the request is also clearly for a *different album*.	Talking Heads track (`2VNfJpwdEQBLyXajaa6LWT`); assistant recommends *Burning Down the House*.	**Scored.** First formally evaluated turn. We score it with $C(q_3) = \{c_{\text{era}}, c_{\text{style}}\}$.
4	The user now makes the exclusion explicit: *not Talking Heads*, while repeating *alternative/industrial rock* and *late 1990s*.	Talking Heads track (`7CqleiaEqHVazV19P532X9`); system again returns a Talking Heads item.	**Scored.** Second evaluated turn. We score it with $C(q_4) = \{c_{\text{avoid}}, c_{\text{era}}, c_{\text{style}}\}$.
5	The user repeats the same constraints once more, emphasizing that the target is a provocative late–1990s alternative/industrial album with a red–haired androgynous figure.	Marilyn Manson track (`1KQxH1Z1BiSo3MMukVpRfl`); the system finally switches away from Talking Heads and returns *The Dope Show*.	**Scored.** Third evaluated turn. We score it with the same constraint set as turn 4, $C(q_5) = \{c_{\text{avoid}}, c_{\text{era}}, c_{\text{style}}\}$.
6	The user explicitly confirms success: *"That's the one! Marilyn Manson's 'Mechanical Animals'!"*	System continues with another track from the correct album neighborhood.	Post–success continuation. This turn confirms that the identification goal has been achieved, but it is not part of the present G1 scoring window.

3.2.3 Running Examples. The universal metrics are most informative when they are read as a coordinated diagnostic set rather than as isolated scores. The following two case studies instantiate the proposed framework on concrete conversations from TalkPlayData–2 [44]. The first example focuses on G1 and shows that catalog grounding does not imply request satisfaction. The second example focuses on G2 and G3 and shows that measurable personalization gain does not necessarily imply discovery. Table 7 summarizes the dataset fields used in the analysis, and Table 8 summarizes the turn structure of the first session.

Running Example 1: Per–turn G1 evaluation on a TalkPlayData–2 session. Session `12286__2016-02-17` involves a listener from Belarus who is trying to identify a specific album from a partial memory of its cover art. The conversational goal is thus not generic recommendation, but constraint refinement over time until one specific target can be recovered. This makes the session especially suitable for illustrating the distinction between *groundedness* and *query adherence*: the system may keep returning real catalog items while still failing to satisfy the actual request. An overview of the session is presented below.

Session `12286__2016-02-17`: user, goal, and request evolution

User profile. A 25-year-old male from Belarus (`country_code = BY`) with English as preferred language and a stated interest in Western alternative rock.

Conversation goal. Category C, high specificity: the listener wants to identify a single album based on its "distinctive and provocative cover art featuring an androgynous figure".

Why this session is useful. The user begins with a vague cover description and gradually introduces precise metadata constraints (genre, era, exclusion). Meanwhile, the system repeatedly suggests Talking Heads until the final scored turn. This illustrates how an LLM can be both grounded and non-adherent at intermediate turns.

Turn progression.

- **Turns 1–2 (context):** broad description and clarification that the task is album identification, not favourite-artist recommendation.
- **Turn 3:** new constraints "alternative/industrial rock" and "late 1990s."
- **Turn 4:** adds the exclusion "not Talking Heads."
- **Turn 5:** repeats all constraints; the system finally recommends Marilyn Manson.
- **Turn 6:** user confirms *Mechanical Animals*.

Goal–progress annotations. The early turns are annotated as not moving toward the goal, which is consistent with the repeated Talking Heads recommendations and with the user's growing frustration [30].

Metric computation. In this example, we compute the G1 metrics only for turns in which the user request can be translated into explicit, checkable constraints. Turns 1–2 are therefore used as conversational context, but are not included in the quantitative metric computation. At that stage, the user only describes a remembered album cover, and the request does not yet specify stable catalog-level constraints such as genre, era, artist exclusion, or release type.

The first turn included in the metric computation is Turn 3. At this point, the user introduces two constraints that can be operationalized using catalog metadata: an era constraint and a style constraint. We thus define:

$$C(q_3) = \{c_{\text{era}}, c_{\text{style}}\}, \qquad c_{\text{era}} : \text{year} \in [1996, 1999], \qquad c_{\text{style}} : \text{alternative/industrial rock}.$$

Subsequent turns are evaluated in the same way, using the updated constraint set implied by the latest user request. For instance, when the user explicitly indicates that the target is not Talking Heads, an additional exclusion constraint is added to the constraint set.

At Turn 4, the exclusion constraint becomes explicit:

$$C(q_4) = \{c_{\text{avoid}}, c_{\text{era}}, c_{\text{style}}\}, \qquad c_{\text{avoid}} : \text{artist} \neq \text{Talking Heads}.$$

Turn 5 keeps the same scored constraint set:

$$C(q_5) = \{c_{\text{avoid}}, c_{\text{era}}, c_{\text{style}}\}.$$

Turn-level results. At Turns 3 and 4, the returned items resolve correctly to catalog entities, so:

$$\text{EntityGroundedness@1}(q_3) = 1, \text{EntityGroundedness@1}(q_4) = 1$$

$$\text{HallucinationRate@1}(q_3) = 0, \text{HallucinationRate@1}(q_4) = 0$$

However, the recommendations remain non-adherent:

$$\text{QueryAdherence@1}(q_3) = 0, \text{QueryAlignment@1}(q_3) = 0, \text{ConstraintHitRate@1}(q_3) = 0.$$

$$\text{QueryAdherence@1}(q_4) = 0, \text{QueryAlignment@1}(q_4) = 0, \text{ConstraintHitRate@1}(q_4) = 0.$$

This means that the system is fully grounded and simultaneously fully wrong with respect to the operative request. If the explanation text names the same returned Talking Heads track, then $\text{ExplanationGroundedness}(q_3) = \text{ExplanationGroundedness}(q_4) = 1$, which is itself informative: explanation consistency does not rescue recommendation failure when the wrong entity is being explained faithfully.

At Turn 5, the system finally switches to the correct catalog neighbourhood and all G1 scores become maximal:

$$\text{EntityGroundedness@1}(q_5) = 1, \qquad \text{HallucinationRate@1}(q_5) = 0,$$

$$\text{QueryAdherence@1}(q_5) = 1, \text{QueryAlignment@1}(q_5) = 1, \text{ConstraintHitRate@1}(q_5) = 1.$$

Session-level summary. Averaging over the scored turns $t \in \{3, 4, 5\}$ gives

$$\overline{\text{EntityGroundedness@1}} = 1, \qquad \overline{\text{HallucinationRate@1}} = 0,$$

$$\overline{\text{QueryAdherence@1}} = \frac{1}{3}, \qquad \overline{\text{QueryAlignment@1}} = \frac{1}{3}, \qquad \overline{\text{ConstraintHitRate@1}} = \frac{1}{3}.$$

Interpretation. This example isolates the core rationale for reporting multiple G1 metrics rather than a single validity score. Catalog grounding would make the system appear strong, because every scored recommendation is executable and non-hallucinated. The adherence metrics reveal a different story: until the final turn, the model repeatedly fails to follow the actual request. The case thus show that *short-term satisfaction* depends on both real-item grounding and faithful constraint following, and that the two should not be conflated.

Running Example 2: G2–G3 evaluation on an exploratory query. Session `53726__2011-12-26` involves a listener from Brazil who explicitly asks to discover *new* artists with an intense or dramatic post-hardcore vibe. The conversation is thus discovery-oriented. This makes the session appropriate for separating G2 and G3: the recommender may still exploit prior taste information successfully while failing the user's current novelty goal. The session overview is presented below.

Session `53726__2011-12-26`: user, goal, and request evolution

User profile. A 20-year-old female from Brazil whose listening history is dominated by Western alternative rock.
Conversation goal. The listener states she wants to discover "new artists" with an intense or dramatic post-hardcore vibe.
Why this session is useful. This conversation highlights how novelty requests can be misinterpreted by a recommender. The user begins by asking to *"discover some new artists"* and specifies a desired mood (*intense or dramatic*). Because a previous session involved an Alesana track, the system continues to recommend Alesana songs: its first response notes that it is *"sticking with"* that band because the listener enjoyed it. Later, the user strengthens the novelty constraint (e.g., *"by an artist I haven't heard before"* or *"not Alesana"*), yet the system again suggests Alesana. This mismatch between user intent and system behavior makes the session a concrete example of a recommendation failure.
Why this session is useful. The system repeatedly suggests Alesana, even after the user explicitly excludes that artist. This exposes a divergence between personalization (reusing prior taste) and discovery (seeking novelty).
Turn progression.

- **Turn 1:** user asks for new, intense music; the system plays Alesana.
- **Turn 2:** reiterates novelty and vibe; the system plays Alesana again.
- **Turn 3:** explicitly "not Alesana"; the system still returns Alesana.

- **Subsequent turns:** the user continues to seek new artists that fit the requested vibe; however, goal progress annotations for the early turns mark them as *not moving toward the goal*, reflecting the recommender's inability to adjust to the novelty constraint.

Goal–progress annotations. According to the dataset's goal progress assessments, the first turn is unlabelled, the second turn moves toward the goal, but turns three and four are annotated as *does_not_move_toward_goal*. These annotations align with the user's growing frustration at being recommended music from the same artist despite asking for new artists.

Scoring setup. The scored turn is Turn 3, where the request can be operationalized as

$$q_3 = \text{“something by an artist I have not heard before, with intense post-hardcore energy”},$$

with:

$$C(q_3) = \{c_{\text{novel}}, c_{\text{style}}, c_{\text{vibe}}\},$$

where,

$$c_{\text{novel}} : \text{artist} \neq \text{Alesana}, \qquad c_{\text{style}} : \text{post-hardcore}, \qquad c_{\text{vibe}} : \text{intense/dramatic}.$$

To illustrate how the list-level metrics are computed, suppose that the recommender returns the following top-3 list for query q_3: $R_K(q_3, u) = (i_1, i_2, i_3)$. In this illustrative case, assume that all three recommended items are valid Alesana tracks. The example is not intended to reproduce the full hidden ranking of the system, but to make the metric calculations transparent for the observed failure mode.

Metric values.

Because all three recommended items are valid catalog entries,

$$\text{EntityGroundedness@3}(q_3) = 1, \qquad \text{HallucinationRate@3}(q_3) = 0.$$

Let's assume that each returned track satisfies the style and vibe requirements but violates the novelty requirement, i.e., the artist is still Alesana. If novelty is treated as a hard constraint, then no item satisfies the full hard constraint set:

$$\text{QueryAdherence@3}(q_3) = 0, \qquad \text{ConstraintHitRate@3}(q_3) = 0.$$

By contrast, soft query alignment can still assign partial credit. With equal weights over the three requirements $\{c_{\text{novel}}, c_{\text{style}}, c_{\text{vibe}}\}$, each item satisfies two out of three requirements, and thus:

$$\text{QueryAlignment@3}(q_3) = \frac{2}{3}.$$

Thus, the list partially matches the musical description of the request, since it captures the desired style and vibe, but it fails the explicit novelty requirement. This illustrates why strict adherence and soft alignment should be reported separately.

From the G2 perspective, assume that all three Alesana tracks belong to the popularity head and are near-redundant under the selected dissimilarity function. Then:

$$\text{LongTailShare@3} = 0, \qquad \text{ARP@3} = 1, \qquad \text{ILD@3} = 0.$$

For a singleton evaluation set $Q = \{q_3\}$, catalog coverage is:

$$\text{CatalogCoverage@3}(Q) = \frac{3}{|I|},$$

which is formally correct, but not substantively informative on its own. From the G3 perspective, let us define the utility U of the items in the recommendation list as the mean satisfaction of the three facets $\{c_{\text{novel}}, c_{\text{style}}, c_{\text{vibe}}\}$. Since each Alesana track satisfies two of the three, the full-system utility is:

$$U(\mathrm{R}(q_3, u)) = \frac{2}{3}.$$

If a profile-ablated baseline satisfies none of the three facets, then

$$U(\mathrm{R}(q_3, \emptyset)) = 0, \qquad \Delta_{\text{profile}} = \frac{2}{3}.$$

At the same time, because the returned items belong to an already-known artist,

$$\text{UserNovelty@3}(q_3, u) = 0.$$

This means that the recommender extracts useful information from the user profile, but uses it to reinforce familiarity rather than support discovery. Even if InterUserOverlap were low against a very different user, that would not change the central point: *inter-user differentiation* and *within-user novelty* are distinct properties.

Interpretation. This case shows why G2 and G3 should not be collapsed into a single notion of recommendation quality. The system exhibits positive personalization gain, because conditioning on the user profile improves alignment with the desired post-hardcore style and mood. Yet it fails the discovery goal completely, because the returned list remains head-heavy, redundant, and non-novel for the current user. The example therefore illustrates a central tension in music recommendation: a system can personalize successfully in the sense of exploiting past taste, while still failing to create *long-term value* through exploration.

3.2.4 Setting-specific Diagnostics. In the following, we discuss setting-specific diagnostic metric families for *few-shot in-context learning (ICL)*, *retrieval-augmented generation (RAG)*, and *chain-/tree-of-thought (CoT/ToT)*. By setting-specific diagnostics, we mean metrics that are not universal to all LLM-based music recommenders, but are useful for analyzing failure modes introduced by a particular configuration. For example, ICL requires diagnostics for prompt sensitivity and example selection, RAG requires diagnostics for evidence grounding and retrieval freshness, and CoT/ToT requires diagnostics for reasoning faithfulness and sequence coherence. Table 6 provides a concise summary and reference frame for these diagnostics in NL-based evaluation of LLM-driven MRSs.

Few-shot In-Context Learning (S-ICL). The evaluation of few-shot ICL might make use of several specialized metrics to assess the *stability* and *utility* of few-shot prompting in music recommendation. For instance, *shot calibration* [214] tests how performance changes as the number of examples (shots) varies; ideally, adding more examples should not drastically mislead the model. Zhang et al. [214] show that as we increase the number of in-context examples, models often become more miscalibrated before improving, with particularly poor calibration in low-shot regimes, highlighting the need to monitor this trade-off. The authors suggest measuring calibration curves across different shot counts (e.g., 0, 3, 5, 7) to ensure the model benefits from additional examples. Zhou et al. [222] propose batch calibration to address the contextual bias that arises from varying numbers of in-context examples, showing that proper calibration techniques can recover performance degradation caused by prompt brittleness across different shot configurations.

Stability to prompt perturbations is another critical dimension: the model's recommendations should remain consistent under minor prompt modifications (e.g., reordering or paraphrasing the examples). For instance, Zhuo et al. [224] introduce PromptSensiScore (PSS), which quantifies prompt sensitivity as the average discrepancy in LLM responses when confronted with semantic variants of the same instruction. Their findings reveal that few-shot examples can

alleviate sensitivity, with the most pronounced reduction occurring in the transition from zero-shot to one-shot settings. For measuring list-based recommendation stability, rank-biased overlap [197] provides a complementary assessment of ranking consistency under perturbations.

Additionally, *preference sensitivity* could be evaluated to ensure that when user preference cues (likes/dislikes) are given as examples, the recommender's output shifts appropriately along the intended facets (such as recommending more of a liked genre and fewer disliked styles). For instance, if a user's example indicates love of jazz and dislike of loud rock, the subsequent recommendations should reflect those biases measurably. Yu et al. [204] propose In-Context Preference Learning (ICPL), showing that LLMs can learn reward functions from iteratively placed human feedback in context, achieving orders of magnitude greater sample efficiency than traditional preference-based methods. Their work shows that properly designed ICL can capture preference signals and translate them into consistent recommendation adjustments, validating that LLMs possess native preference-learning capabilities that enable few-shot adaptation to user tastes.

Finally, *instruction–example alignment* might check that the system continues to obey the explicit request instructions even when some in-context examples might be ambiguous or in conflict with those instructions. In other words, the user's prompt requirements (e.g., "no mainstream hits") must dominate over any contradicting signals from the examples (e.g., in case these include predominantly chart music). Zhou et al. [222] show that LLM outputs can be sensitive to seemingly minor prompt-design choices, including formatting, label wording in classification-style prompts, and the selection of in-context examples. Such factors can cause the model to deviate from the intended instruction. Zhuo et al. [224] further show that instruction-following fidelity is related to model confidence, with higher-confidence generations being more robust to prompt variations. In the context of ICL-based music recommendation, these findings motivate robustness checks that vary prompt templates, example order, and example selection.

Retrieval-Augmented Generation (S-RAG). The evaluation of an S-RAG pipeline spans multiple aspects across both the retrieval and recommendation stages. As noted by Es et al. [50], evaluating RAG systems requires assessing the retrieval module's ability to fetch relevant context and the generation module's ability to use that evidence. For the retrieval stage, one can use standard IR metrics (e.g., Recall@K of known relevant items) to gauge if the LLM is retrieving useful evidence. At the recommendation generation stage, the primary focus is on *grounding*: conditioning the LLM on up-to-date catalog facts should reduce hallucinations and ensure that recommended tracks exist and meet the user's criteria. To quantify groundedness, one might define *EvidenceGrounding@K* as the proportion of the top-K recommended items whose key attributes can be justified by the retrieved items. This metric is inspired by faithfulness and citation-based evaluations proposed for RAG outputs [8, 50].

Another diagnostic is a document-swap test, denoted Δdoc, where we perturb the retrieved context (e.g., replacing a percentage of relevant items with irrelevant ones) to see if the model's recommendations change. A strong drop in recommendation quality after such a swap would indicate that the model truly relies on the retrieved evidence, whereas an invariant output would suggest that it ignores the provided items. This methodology is analogous to input-perturbation analyses used to study how LLMs utilize long contexts [107].

Beyond grounding, we also need to evaluate the *freshness*, *breadth*, and *efficiency* of RAG-based recommendations. For timeliness, inspired by work on time-aware evaluation and recency effects in RSs [18, 34], one might introduce freshness and temporal-bias metrics such as *Time-To-First-Surface (TTFS)* and *Day-0 Lift*, whose aim is to summarize how quickly new releases first appear in either the retrieved pool or the final top-K list, and how much extra exposure

they receive on their release day. For example, a low TTFS and high Day-0 Lift would indicate that a newly released track already shows up quickly and prominently in users' recommendations instead of only weeks later.

To understand which parts of the catalog the LLM actually uses, we monitor simple coverage-style metrics inspired by beyond-accuracy evaluation [89, 174], notably metrics such as *UniqueArtists@K* counting distinct artists in the top-K recommendations, or *EntitiesResolvedShare* measuring the fraction of recommended items that can be matched to real catalog entities (i.e., non-hallucinated tracks).

Finally, it is important to monitor the inference latency L (end-to-end response time), the total token budget B (number of tokens processed per query, including prompt, retrieved evidence, and generated output), and the evidence token ratio B_D/B (share of the budget devoted specifically to retrieved documents). An effective S-RAG configuration should improve freshness and catalog coverage at both retrieval and recommendation stages, without incurring prohibitive cost, while still satisfying core recommendation goals such as accuracy, novelty, and personalization.

Chain-of-Thought Reasoning (S-COT). Complex reasoning for playlist curation requires both verifiable reasoning steps and coherent output. Here, one might use *Reasoning Faithfulness*, which measures intermediate steps verifiable against catalogs [137]. In contrast, *Reasoning Impact* (ΔU_{CoT}) can be used to quantify the utility gain from explicit reasoning chains compared to direct answers [199]. *Self-consistency Uplift* evaluates improvements from aggregating outputs across multiple reasoning paths [195]. Overall, one can state that the reasoning metrics integrate with query fidelity, diversity, and controllability dimensions.

3.2.5 Measures of Risk. Beyond success-oriented metrics, which capture how well a system fulfills user intent and provides relevant, grounded recommendations, it is equally important to account for potential *risks* and undesirable behaviors arising from generative models in music recommendation. For this reason, we include a complementary set of **diagnostic metrics and dimensions** addressing issues such as hallucination, exposure skew, privacy and stereotyping, automatic-judge bias, and cross-lingual failure. These diagnostics provide a broader understanding of system robustness, fairness, and trustworthiness—dimensions that classical success metrics alone cannot capture. We identify the following risks:

R1. Hallucinations and entity errors. This dimension measures instances of fabrication (generating non-existent entities), misattribution (assigning incorrect properties to real entities), or failed resolution against the catalog or an external knowledge base.

R2. Popularity, temporal, and language bias. This dimension evaluates systematic disparities in exposure across stratified groups, such as mainstream vs. long-tail content or Western vs. non-Western music.

R3. Profile hazards. This dimension assesses the system's sensitivity to variations in user profile representation, including robustness to profile edits (addition/removal of preference signals), degradation with truncated profiles (measuring utility loss when profile length is constrained), and the impact of noisy or contradictory preference signals.

R4. Evaluator bias. When LLMs serve as automatic evaluators (LLM-as-judge), this dimension quantifies systematic divergence between LLM-based assessments and human ground-truth judgments.

R5. Sampling-bias amplification. This dimension evaluates how the composition of in-context examples systematically affects exposure distribution for specific groups (e.g., language, genre, or artist popularity tier).

R6. Order brittleness and query contradiction. These criteria quantify recommendation instability arising from perturbations in example ordering or conflicts between explicit instructions and implicit example patterns.

R7. Privacy leakage and reproducibility. These complementary dimensions assess seed example memorization in outputs (privacy risk) and generation determinism (reproducibility).

R8. Implicit-feedback noise. This dimension quantifies how noisy implicit signals, e.g., skips or incomplete plays, in seed examples affect recommendation utility.

R1. Hallucinations and entity errors. Ji et al. [88] provide a comprehensive taxonomy distinguishing between factuality hallucinations (contradicting established knowledge) and faithfulness hallucinations (inconsistency with provided context). For RAG systems, counterfactual robustness testing can be applied: evaluating if the model appropriately handles retrieved documents containing incorrect or contradictory information, or measuring if outputs change when relevant documents are replaced with irrelevant ones [50]. High-quality RAG systems should show substantial sensitivity to document relevance, i.e., changing outputs significantly when retrieved context changes while maintaining the ability to reject answers when retrieved documents lack relevant information. In MRSs, this can be quantified as the proportion of recommended tracks that do not exist in the catalog or have incorrectly attributed metadata such as artist, album, or genre.

R2. Popularity, temporal, and language bias. Abdollahpouri et al. [3] introduce the Average Recommendation Popularity (ARP) metric: $\text{ARP} = \frac{1}{|U|} \sum_{u \in U} \frac{\sum_{i \in R_u} \phi(i)}{|R_u|}$, where $\phi(i)$ represents item i's popularity score and R_u is the recommendation list for user u. They show that collaborative filtering algorithms exhibit systematic bias toward popular items, with ARP values significantly exceeding catalog median popularity. LLMs may exacerbate this bias or introduce new types of bias. Deldjoo et al. [34] show that ChatGPT-based RSs exhibit provider fairness issues, temporal instability, and recency bias in music and movie recommendations, evaluating consistency through repeated queries and temporal drift analysis. Sguerra et al. [170] reveal that LLM-generated musical taste profiles exhibit systematic quality variations across user mainstreamness and cultural diversity, measuring profile accuracy through genre precision and user study evaluations.

R3. Profile hazards. Wang et al. [191] introduce profile defect diagnostics, categorizing issues as inaccurate (contradictory preferences), incomplete (missing critical attributes), or both. Evaluation involves measuring recommendation stability (via Rank-Biased Overlap [197]) across profile perturbations and quantifying the accuracy-length trade-off: plotting recommendation quality (nDCG or Recall) against profile token length to identify optimal compression points. For LLM-based systems, this includes assessing if prompt length constraints (e.g., 2048 vs. 8192 tokens) significantly degrade personalization quality.

R4. Evaluator bias. Zheng et al. [221] show that LLM judges exhibit multiple forms of bias: position bias (where GPT-4 shows only 65% consistency when answer positions are swapped, favoring the first position in 30% of cases), verbosity bias (where judges may prefer longer responses, with GPT-4 showing 8.7% susceptibility to verbosity attacks compared to over 90% for weaker models), and self-enhancement bias (where GPT-4 exhibits approximately 10% higher win rates for its own outputs compared to human judgments). To measure evaluator bias, Zheng et al. propose conducting position swap tests to assess consistency and comparing agreement rates between LLM-human pairs versus human-human

pairs, finding that GPT-4 achieves over 80% agreement with human preferences when biases are mitigated. Additionally, Wang et al. [80] provide comprehensive analysis of LLM-as-judge reliability across diverse evaluation scenarios.

R5. Sampling-bias amplification. Lin et al. [104] show that recommendation models memorize item popularity in the principal spectrum and amplify this bias. For ICL contexts, bias amplification can be quantified by comparing group representation in recommendations versus catalog baselines, measured as relative popularity lift: $\Delta\text{Pop}(g) = \frac{\text{Pop}_{\text{rec}}(g) - \text{Pop}_{\text{catalog}}(g)}{\text{Pop}_{\text{catalog}}(g)}$, where positive values indicate over-representation. Similarly, user-level popularity calibration [98] identifies popularity mismatches between historic user interactions H_u and recommended items R_u, measured as Jensen-Shannon Divergence (JSD) between the popularity distribution of items in H_u and R_u (over a fixed number of popularity bins). Evaluation should stratify across demographic dimensions and employ counterfactual seed sets to isolate causal effects of example composition.

R6. Order brittleness and query contradiction. PromptSensiScore [224] measures sensitivity to prompt variants, finding that few-shot examples can alleviate instability with the most pronounced reduction in zero-to-one-shot transitions. Order brittleness can be measured via rank-biased overlap [197] across different orderings: $\text{OB} = 1 - \text{RBO}(R_{\pi_1}, R_{\pi_2})$, where π_1, π_2 represent different orderings of identical examples. Query contradiction is assessed by providing examples that violate stated constraints and measuring constraint adherence. When substantial instability is detected, fallback strategies include zero-shot prompting or explicit alignment scaffolds emphasizing instruction priority over example patterns.

R7. Privacy leakage and reproducibility. Carlini et al. [20] show that LLMs memorize training data, which can even be extracted in terms of verbatim sequences, particularly for repeated or distinctive content. In ICL contexts, leakage is measured as the proportion of seed example text surfacing in generated rationales or recommendations. For reproducibility, seed sets should be cryptographically hashed and logged alongside model version, temperature (with $T = 0$ ensuring determinism), and sampling parameters. Reproducibility enables systematic auditing when recommendation quality degrades, verified by regenerating recommendations with identical inputs and measuring output consistency.

R8. Implicit-feedback noise. Hu et al. [79] show that treating all implicit feedback uniformly degrades collaborative filtering performance, as signals like skips carry inherent ambiguity (dislike versus context inappropriateness). For ICL evaluation, one might compare recommendation quality across seed configurations: all interactions, positive-only signals, and confidence-weighted signals (e.g., high play completion rates), measuring utility via ranking metrics (nDCG@K) and diversity across configurations. When substantial quality degradation occurs with noisy seeds, filtering strategies should weight implicit signals by confidence indicators or exclude ambiguous interactions from ICL contexts.

3.2.6 Further Considerations on User and Item Modeling. Since the goal of automatically generated user profiles or item descriptions is to replace traditional user or item representation in a recommendation setting, the most straightaway methods of evaluating their success is through traditional recommendation metrics such as Recall, and NDCG. In production, if users or content editors are allowed to edit user profiles or music content descriptors, respectively, then the amount of correction required serves as a good indicator of the approach's ability to accurately represent users or items. However, most of the work on using LLMs to summarize users' preferences in a scrutable way has focused on domains such as books, movies, or e-commerce [64, 148, 223], whereas music presents its own unique challenges.

Given the context-length limitations of LLMs, it becomes crucial to devise effective sampling strategies for selecting items that best capture users' interests. Moreover, in the absence of explicit textual feedback, user or item modeling with LLMs must rely on item metadata, whose quality can vary substantially [120]. Also, various kinds of biases are present. In [170], authors show that profile quality, split into two dimensions (1) self-identification and (2) downstream recommendation, is biased towards certain types of music genres. For example, the tested LLMs generated better descriptions of taste when the users consumed metal music that tends to be of American origin and is largely discussed on internet forums and magazines. More recent music tracks and content such as french Rap tended to be less well attended, indicating that due to the corpora used to train LLMs, some content types end up being underrepresented. These differences raise fairness questions, e.g., if some users do not perceive their tastes accurately reflected by the system, they are less likely to trust the system. In addition, some items might be less reliably described by an LLM.

Moreover, due to the dynamic nature of conversations in LLMs, established datasets for offline evaluation are generally insufficient. Instead, users' explicit indication of their queries, emotions, intents, etc. need to be collected. Finally, assessing the contribution of advanced user models that integrate personas or psychological factors with music preferences in recommendation is considerably challenging, as disentangling the underlying factors is a non-trivial task.

LLMs and Long-Tail Items. LLMs struggle with *long-tail entities* [71, 82], a challenge that is particularly present in the music industry, where a large portion of the catalog consists of such lesser-known music artists or tracks. Recent research [82] has explored ways to address this limitation by introducing non-parametric knowledge into LLMs—either by augmenting prompts with relevant external information about the items or by injecting structured triplets from knowledge bases. The authors' findings show that these approaches can significantly improve performance. In particular, prompting LLMs with knowledge base triples tends to work better than just injecting external information about the items in the prompt, while combining both methods can substantially reduce hallucinations in the generated content. However, when long-tail items correspond to new tracks or artists for which such external information is not yet available, either textually or within knowledge bases, this strategy remains infeasible, and alternative approaches still need to be explored. Evaluating performance on long-tail items is thus crucial, and at the very least, their impact should be systematically monitored.

LLMs and Personas. Personalization lies at the core of RSs. In the context of LLMs, it could refer to their ability to represent different users or user groups, i.e., to go beyond an "average" listener or music and adapt to diverse music preferences and behaviors. In other words, it reflects the model's steerability: its capacity to adjust its outputs based on the user. In practice, steerability can be achieved through prompting, fine-tuning, or activation steering [127]. Prompting remains the most common approach, as fine-tuning and activation methods often require access to model internals or significant computational resources. However, Miehling et al. [127] show that prompt-based steerability remains highly challenging: it depends on model size, with larger models generally showing better steerability, and, more importantly, each model tends to be more responsive to specific subsets of persona dimensions by default. Given these findings, how can we fairly assess an LLM's ability to adapt while keeping in mind their inherent bias toward certain dimensions? Future research should measure not only overall quality but also how consistently models reflect individuals or groups, when used either in solutions or as evaluators.

4 Conclusion

In summary, our work highlights how LLMs are already reshaping the development of MRSs. Their ability to interpret and generate text enables richer user and item representations and supports interaction modalities that extend beyond

traditional recommendation pipelines. At the same time, we reiterate the importance of carefully adapting evaluation practices to these emerging capabilities. The generative, context-dependent nature of LLM-based systems challenges long-standing assumptions in MRS evaluation, reinforcing the need for methodologies that remain reliable, transparent, and aligned with real user behavior. As LLMs continue to evolve, ensuring that evaluation frameworks evolve alongside them will be essential for maintaining rigor and trustworthiness in RS research.

We posit that this is an opportune moment for the community to take a step back and reflect on evaluation practices that better capture the underlying goals of music recommendation. Beyond "solving" datasets, we argue for frameworks that address what constitutes a good music recommendation and how systems can support long-term user preferences, fairness, and healthier consumption patterns. Given the nature of LLMs, traditional performance metrics based on predicting consumed items become increasingly insufficient, both because of the opaque provenance of training data, which may overlap with popular recommendation datasets, and because such models can rely heavily on memorization to achieve high item-prediction scores. As we have described, the NLP community is comparatively more advanced in addressing these methodological concerns, and its practices can thus serve as a foundation for future MRSs and general RSs research to build upon.

We highly encourage further research on LLM-enhanced MRSs, to remain attentive to the risks these models introduce, and to design evaluation measures that reflect the goals and complexities of music recommendation. We expect our work to offer a grounded framework to navigate the opportunities, but also challenges of LLM usage, and the inherent subtleties of musical preferences and interactions that require careful consideration when building LLM-based MRSs.

Acknowledgments

This research was funded in whole or in part by the Austrian Science Fund (FWF): 10.55776/COE12, and 10.55776/DFH23, and 10.55776/P36413; and by the State of Upper Austria and the Federal Ministry of Education, Science, and Research through grant LIT-2024-13-SEE-111.

References

[1] Marah I Abdin, Suriya Gunasekar, Varun Chandrasekaran, Jerry Li, Mert Yuksekgonul, Rahee Ghosh Peshawaria, Ranjita Naik, and Besmira Nushi. 2024. KITAB: Evaluating LLMs on Constraint Satisfaction for Information Retrieval. In *The Twelfth International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=b3kDP3IytM

[2] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. 2020. Multistakeholder recommendation: Survey and research directions: H. Abdollahpouri et al. *User Modeling and User-Adapted Interaction* 30, 1 (2020), 127–158.

[3] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. Managing Popularity Bias in Recommender Systems with Personalized Re-ranking. In *Proceedings of the 29th International Florida Artificial Intelligence Research Society Conference (FLAIRS)*. 413–418.

[4] Darius Afchar, Alessandro Melchiorre, Markus Schedl, Romain Hennequin, Elena Epure, and Manuel Moussallam. 2022. Explainability in music recommender systems. *AI Magazine* 43, 2 (2022), 190–208.

[5] Khetam Al Sharou, Zhenhao Li, and Lucia Specia. 2021. Towards a Better Understanding of Noise in Natural Language Processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, Ruslan Mitkov and Galia Angelova (Eds.). INCOMA Ltd., Held Online, 53–62. https://aclanthology.org/2021.ranlp-1.7/

[6] Geetha Sai Aluri, Siddharth Sharma, Tarun Sharma, and Joaquin Delgado. 2024. Playlist Search Reinvented: LLMs Behind the Curtain. In *Proceedings of the 18th ACM Conference on Recommender Systems* (Bari, Italy) *(RecSys '24)*. Association for Computing Machinery, New York, NY, USA, 813–815. doi:10.1145/3640457.3688047

[7] Enrique Amigó, Yashar Deldjoo, Stefano Mizzaro, and Alejandro Bellogín. 2023. A unifying and general account of fairness measurement in recommender systems. *Information Processing & Management* 60, 1 (2023), 103115.

[8] Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, Vol. 2024. 9112–9141.

[9] Yuyang Bai, Shangbin Feng, Vidhisha Balachandran, Zhaoxuan Tan, Shiqi Lou, Tianxing He, and Yulia Tsvetkov. 2024. KGQuiz: Evaluating the Generalization of Encoded Knowledge in Large Language Models. In *Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW*

'24). Association for Computing Machinery, New York, NY, USA, 2226–2237. doi:10.1145/3589334.3645623

[10] Ashmi Banerjee, Adithi Satish, Fitri Nur Aisyah, Wolfgang Wörndl, and Yashar Deldjoo. 2026. Collab-REC: An LLM-based Agentic Framework for Balancing Recommendations in Tourism. *ACM Transactions on Recommender Systems* (2026).

[11] Keqin Bao, Ming Yan, Yang Zhang, Jizhi Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2024. Real-Time Personalization for LLM-based Recommendation with Customized In-Context Learning. arXiv:2410.23136 [cs.IR] https://arxiv.org/abs/2410.23136

[12] Christine Bauer and Markus Schedl. 2019. Global and country-specific mainstreaminess measures: Definitions, analysis, and usage for improving personalized music recommendation systems. *PLOS ONE* 14, 6 (June 2019), e0217389. doi:10.1371/journal.pone.0217389 Publisher: Public Library of Science.

[13] Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 238–255. doi:10.18653/v1/2025.acl-short.20

[14] Giovanni Maria Biancofiore, Tommaso Di Noia, Eugenio Di Sciascio, Fedelucio Narducci, and Paolo Pastore. 2022. Aspect based sentiment analysis in music: a case study with spotify. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing* (Virtual Event) *(SAC '22)*. Association for Computing Machinery, New York, NY, USA, 696–703. doi:10.1145/3477314.3507092

[15] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*. 405–414.

[16] Tony Brooke. 2014. descriptive metadata in the music industry: Why it is broken and how to fix it—part one. *Journal of Digital Media Management* 2, 3 (2014), 263–282.

[17] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS 2020)*. https://arxiv.org/abs/2005.14165 arXiv preprint arXiv:2005.14165.

[18] Pedro G Campos, Fernando Díez, and Iván Cantador. 2014. Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols. *User Modeling and User-Adapted Interaction* 24, 1 (2014), 67–119.

[19] Estefanía Cano, Fernando Mora-Ángel, Gustavo A López Gil, José Ricardo Zapata, Antonio Escamilla, Juan Fernando Alzate, and Moisés Betancur. 2021. Sesquialtera in the Colombian bambuco: Perception and estimation of beat and meter–extended version. *Transactions of the International Society for Music Information Retrieval* 4, 1 (2021).

[20] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*. 2633–2650.

[21] Brandon James Carone and Pablo Ripollés. 2024. SoundSignature: What Type of Music do you Like?. In *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 1–10.

[22] Elif Celen, Pol van Rijn, Harin Lee, and Nori Jacoby. 2025. Are Expressions for Music Emotions the Same Across Cultures?. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*.

[23] Òscar Celma Herrada et al. 2009. *Music recommendation and discovery in the long tail*. Universitat Pompeu Fabra.

[24] Ching-Wei Chen, Paul Lamere, Markus Schedl, and Hamed Zamani. 2018. Recsys challenge 2018: Automatic music playlist continuation. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 527–528.

[25] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems* 41, 3 (2023), 1–39.

[26] Li Chen and Pearl Pu. 2012. Critiquing-based recommenders: survey and emerging trends. *User Modeling and User-Adapted Interaction* 22, 1 (2012), 125–150.

[27] Yanran Chen and Steffen Eger. 2023. MENLI: Robust Evaluation Metrics from Natural Language Inference. *Transactions of the Association for Computational Linguistics* 11 (2023), 804–825. doi:10.1162/tacl_a_00576

[28] Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the Use of Large Language Models for Reference-Free Text Quality Evaluation: An Empirical Study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (Eds.). Association for Computational Linguistics, Nusa Dua, Bali, 361–374. doi:10.18653/v1/2023.findings-ijcnlp.32

[29] Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations?. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 15607–15631. doi:10.18653/v1/2023.acl-long.870

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

[31] Pierre Colombo, Guillaume Staerman, Chloé Clavel, and Pablo Piantanida. 2021. Automatic Text Evaluation through the Lens of Wasserstein Barycenters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 10450–10466. doi:10.18653/v1/2021.emnlp-main.817

[32] Paolo Cremonesi and Dietmar Jannach. 2021. Progress in recommender systems research: Crisis? What crisis? *AI Magazine* 42, 3 (2021), 43–54.

[33] Mathieu Delcluze, Antoine Khoury, Clémence Vast, Valerio Arnaudo, Léa Briand, Walid Bendada, and Thomas Bouabça. 2025. Text2Playlist: Generating Personalized Playlists from Text on Deezer. In *The 47th European Conference on Information Retrieval (ECIR 2025)*.

[34] Yashar Deldjoo. 2026. Understanding Biases in ChatGPT-based Recommender Systems: Provider Fairness, Temporal Stability, and Recency. *Trans. Recomm. Syst.* 4, 2 (2026), 17:1–17:35. doi:10.1145/3690655

[35] Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, Rene Vidal, Maheswaran Sathiamoorthy, Atoosa Kasrizadeh, Silvia Milano, et al. 2026. Recommendation with Generative Models. *Foundations and Trends® in Information Retrieval* 20, 1-2 (2026), 1–176.

[36] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2024. Fairness in recommender systems: research landscape and future directions. *User Modeling and User-Adapted Interaction* 34, 1 (2024), 59–108.

[37] Yashar Deldjoo, Nikhil Mehta, Maheswaran Sathiamoorthy, Shuai Zhang, Pablo Castells, and Julian McAuley. 2025. Toward Holistic Evaluation of Recommender Systems Powered by Generative Models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3932–3942.

[38] Yashar Deldjoo, Markus Schedl, and Peter Knees. 2024. Content-driven music recommendation: Evolution, state of the art, and challenges. *Computer Science Review* 51 (2024), 100618. doi:10.1016/j.cosrev.2024.100618

[39] Zihao Deng, Yinghao Ma, Yudong Liu, Rongchen Guo, Ge Zhang, Wenhu Chen, Wenhao Huang, and Emmanouil Benetos. 2024. MusiLingo: Bridging Music and Text with Pre-trained Language Models for Music Captioning and Query Response. In *Findings of the Association for Computational Linguistics: NAACL 2024*. 3643–3655.

[40] Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the Limitations of Reference-Free Evaluations of Generated Text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 10960–10977. doi:10.18653/v1/2022.emnlp-main.753

[41] Dario Di Palma, Felice Antonio Merra, Maurizio Sfilio, Vito Walter Anelli, Fedelucio Narducci, and Tommaso Di Noia. 2025. Do LLMs Memorize Recommendation Datasets? A Preliminary Study on MovieLens-1M. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 2582–2586. doi:10.1145/3726302.3730178

[42] Karlijn Dinnissen and Christine Bauer. 2022. Fairness in music recommender systems: A stakeholder-centered mini review. *Frontiers in big Data* 5 (2022), 913608.

[43] Seungheon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023. LP-MusicCaps: LLM-Based Pseudo Music Captioning. In *Ismir 2023 Hybrid Conference*.

[44] Seungheon Doh, Keunwoo Choi, and Juhan Nam. 2025. TALKPLAY: Multimodal Music Recommendation with Large Language Models. *arXiv preprint arXiv:2502.13713* (2025).

[45] Seungheon Doh, Junwon Lee, and Juhan Nam. 2021. Music Playlist Title Generation: A Machine-Translation Approach. In *Proceedings of the 2nd Workshop on NLP for Music and Spoken Audio (NLP4MusA)*, Sergio Oramas, Elena Epure, Luis Espinosa-Anke, Rosie Jones, Massimo Quadrana, Mohamed Sordo, and Kento Watanabe (Eds.). Association for Computational Linguistics, Online, 27–31. https://aclanthology.org/2021.nlp4musa-1.6/

[46] SeungHeon Doh, Minhee Lee, Dasaem Jeong, and Juhan Nam. 2024. Enriching Music Descriptions with a Finetuned-LLM and Metadata for Text-to-Music Retrieval. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[47] Elena Epure and Romain Hennequin. 2023. A human subject study of named entity recognition in conversational music recommendation queries. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 1281–1296.

[48] Elena V. Epure, Anis Khlif, and Romain Hennequin. 2019. Leveraging knowledge bases and parallel annotations for music genre translation. In *International Society for Music Information Retrieval Conference*.

[49] Elena V. Epure, Guillaume Salha, Manuel Moussallam, and Romain Hennequin. 2020. Modeling the Music Genre Perception across Language-Bound Cultures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 4765–4779. doi:10.18653/v1/2020.emnlp-main.386

[50] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Nikolaos Aletras and Orphee De Clercq (Eds.). Association for Computational Linguistics, St. Julians, Malta, 150–158. doi:10.18653/v1/2024.eacl-demo.16

[51] Alexander Felfernig, Gerhard Friedrich, Dietmar Jannach, and Markus Zanker. 2015. Constraint-based recommender systems. In *Recommender systems handbook*. Springer, 161–190.

[52] Andres Ferraro. 2019. Music cold-start and long-tail recommendation: bias in deep representations. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys '19)*. Association for Computing Machinery, New York, NY, USA, 586–590. doi:10.1145/3298689.3347052

[53] Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain. In *Proceedings of the Sixth Conference on Machine Translation*, Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz (Eds.). Association for Computational Linguistics, Online, 733–774. https://aclanthology.org/2021.wmt-1.73/

[54] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as You Desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 6556–6576. doi:10.18653/v1/2024.naacl-long.365

[55] Yumeng Fu, Junjie Wu, Zhongjie Wang, Meishan Zhang, Lili Shan, Yulin Wu, and Bingquan Liu. 2025. LaERC-S: Improving LLM-based Emotion Recognition in Conversation with Speaker Characteristics. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 6748–6761. https://aclanthology.org/2025.coling-main.451/

[56] Christian Fuentes, Johan Hagberg, and Hans Kjellberg. 2019. Soundtracking: music listening practices in the digital age. *European Journal of Marketing* 53, 3 (2019), 483–503.

[57] Magdalena Fuentes, Lucas S Maia, Martín Rocamora, Luiz WP Biscainho, Hélene C Crayencour, Slim Essid, and Juan P Bello. 2019. Tracking beats and microtiming in Afro-Latin American music using conditional random fields and deep learning. In *ISMIR*.

[58] Nieves Fuentes-Sánchez, Raúl Pastor, Tuomas Eerola, Miguel A. Escrig, and M. Carmen Pastor. 2022. Musical preference but not familiarity influences subjective ratings and psychophysiological correlates of music-induced emotions. *Personality and Individual Differences* 198 (Nov. 2022), 111828. doi:10.1016/j.paid.2022.111828

[59] Giovanni Gabbolini and Derek Bridge. 2021. Generating Interesting Song-to-Song Segues With Dave. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 98–107. doi:10.1145/3450613.3456819

[60] Giovanni Gabbolini, Romain Hennequin, and Elena Epure. 2022. Data-Efficient Playlist Captioning With Musical and Linguistic Knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 11401–11415. doi:10.18653/v1/2022.emnlp-main.784

[61] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten De Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI open* 2 (2021), 100–126.

[62] Qingqing Gao, Jiuxin Cao, Biwei Cao, Xin Guan, and Bo Liu. 2024. CEPT: A Contrast-Enhanced Prompt-Tuning Framework for Emotion Recognition in Conversation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 2947–2957. https://aclanthology.org/2024.lrec-main.263/

[63] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* 2, 1 (2023).

[64] Zhaolin Gao, Joyce Zhou, Yijia Dai, and Thorsten Joachims. 2024. End-to-end Training for Recommendation with Language-based User Profiles. *arXiv preprint arXiv:2410.18870* (2024).

[65] Iker García-Ferrero, Begoña Altuna, Javier Alvez, Itziar Gonzalez-Dios, and German Rigau. 2023. This is not a Dataset: A Large Negation Benchmark to Challenge Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 8596–8615. doi:10.18653/v1/2023.emnlp-main.531

[66] Josh Gardner, Simon Durand, Daniel Stoller, and Rachel Bittner. 2024. LLARK: a multimodal instruction-following language model for music. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 603, 46 pages.

[67] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2025. Scaling Synthetic Data Creation with 1,000,000,000 Personas. doi:10.48550/arXiv.2406.20094 arXiv:2406.20094 [cs].

[68] Daniel Gómez-Marín, Rafael Ospina-Caicedo, Javier Díaz-Cely, Jesús Paz, Sergi Jordà, and Perfecto Herrera. 2024. Salsa, a dataset for beat estimation in salsa music. *Transactions of the International Society for Music Information Retrieval* 7, 1 (2024).

[69] Jiatao Gu, Kyunghyun Cho, and Victor O.K. Li. 2017. Trainable Greedy Decoding for Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 1968–1978. doi:10.18653/v1/D17-1210

[70] Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2024. APPLS: Evaluating Evaluation Metrics for Plain Language Summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 9194–9211. doi:10.18653/v1/2024.emnlp-main.519

[71] Simon Hachmeier and Robert Jäschke. 2025. A Benchmark and Robustness Study of In-Context-Learning with Large Language Models in Music Entity Detection. In *Proceedings of the 31st International Conference on Computational Linguistics*. 9845–9859.

[72] Anna Hausberger, Hannah Strauss, and Markus Schedl. 2025. ExIM: Exploring Intent of Music Listening for Retrieving User-generated Playlists. In *Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '25)*. Association for Computing Machinery, New York, NY, USA, 348–357. doi:10.1145/3698204.3716470

[73] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[74] Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. 2023. Large language models as zero-shot conversational recommenders. In *Proceedings of the 32nd ACM international conference on information and knowledge management*. 720–730.

[75] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)* 22, 1 (2004), 5–53.

[76] Yoshinori Hijikata, Kazuhiro Iwahama, and Shogo Nishida. 2006. Content-based music filtering system with editable user profile. In *Proceedings of the 2006 ACM symposium on Applied computing*. 1050–1057.

[77] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rygGQyrFvH

[78] Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*. Springer, 364–381.

[79] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE international conference on data mining*. Ieee, 263–272.

[80] Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2025. An Empirical Study of LLM-as-a-Judge for LLM Evaluation: Fine-tuned Judge Model is not a General Substitute for GPT-4. arXiv:2403.02839 [cs.CL] https://arxiv.org/abs/2403.02839

[81] Jin Huang, Harrie Oosterhuis, Masoud Mansoury, Herke Van Hoof, and Maarten de Rijke. 2024. Going beyond popularity and positivity bias: Correcting for multifactorial bias in recommender systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 416–426.

[82] Wenyu Huang, Guancheng Zhou, Mirella Lapata, Pavlos Vougiouklis, Sebastien Montella, and Jeff Z. Pan. 2025. Prompting large language models with knowledge graphs for question answering involving long-tail facts. *Knowledge-Based Systems* 324 (Aug. 2025), 113648. doi:10.1016/j.knosys.2025.113648

[83] Karim M Ibrahim, Elena V Epure, Geoffroy Peeters, and Gael Richard. 2020. SHOULD WE CONSIDER THE USERS IN CONTEXTUAL MUSIC AUTO-TAGGING MODELS?. In *21st International Society for Music Information Retrieval Conference*.

[84] Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull Soc Vaudoise Sci Nat* 37 (1901), 547–579.

[85] Ari Jacovi and Yoav Goldberg. 2020. Aligning Faithful Explanations: Towards Explanation as Verification. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)* (2020), 514–525. https://aclanthology.org/2020.emnlp-main.40

[86] Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 4198–4205.

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

[88] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys* 55, 12 (2023), 1–38.

[89] Marius Kaminskas and Derek Bridge. 2016. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 1 (2016), 1–42.

[90] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.

[91] Shah Noor Khan, Eelco Herder, and Diba Kaya. 2024. Experiences of Non-Mainstream and Minority Users with Music Recommendation Systems. In *Mensch und Computer 2024-Workshopband*. Gesellschaft für Informatik eV, 10–18420.

[92] Mohammad Khosravani, Chenyang Huang, and Amine Trabelsi. 2024. Enhancing Argument Summarization: Prioritizing Exhaustiveness in Key Point Generation and Introducing an Automatic Coverage Evaluation Metric. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 8212–8224. doi:10.18653/v1/2024.naacl-long.454

[93] Bart P Knijnenburg and Martijn C Willemsen. 2015. Evaluating recommender systems with user experiments. In *Recommender systems handbook*. Springer, 309–352.

[94] Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio Flamingo: a novel audio language model with few-shot learning and dialogue abilities. In *Proceedings of the 41st International Conference on Machine Learning*. 25125–25148.

[95] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The unfairness of popularity bias in music recommendation: A reproducibility study. In *European conference on information retrieval*. Springer, 35–42.

[96] Matevž Kunaver and Tomaž Požrl. 2017. Diversity in recommender systems–A survey. *Knowledge-based systems* 123 (2017), 154–162.

[97] H. Lee, E. Çelen, P. M. C. Harrison, M. Anglada-Tort, P. van Rijn, M. Park, M. Schönwiesner, and N. Jacoby. 2025. GlobalMood: A Cross-Cultural Benchmark for Music Emotion Recognition. In *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*.

[98] Oleg Lesota, Stefan Brandl, Matthias Wenzel, Alessandro B. Melchiorre, Elisabeth Lex, Navid Rekabsaz, and Markus Schedl. 2022. Exploring Cross-group Discrepancies in Calibrated Popularity for Accuracy/Fairness Trade-off Optimization. In *Proceedings of the 2nd Workshop on Multi-Objective Recommender Systems co-located with 16th ACM Conference on Recommender Systems (RecSys 2022), Seattle, WA, USA, 18th-23rd September 2022*

(CEUR Workshop Proceedings, Vol. 3268), Himan Abdollahpouri, Shaghayegh Sahebi, Mehdi Elahi, Masoud Mansoury, Babak Loni, Zahra Nazari, and Maria Dimakopoulou (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-3268/paper5.pdf

[99] Oleg Lesota, Gustavo Escobedo, Yashar Deldjoo, Bruce Ferwerda, Simone Kopeinik, Elisabeth Lex, Navid Rekabsaz, and Markus Schedl. 2023. Computational Versus Perceived Popularity Miscalibration in Recommender Systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 1889–1893. doi:10.1145/3539618.3591964

[100] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[101] Ang Li, Jennifer Thom, Praveen Chandar, Christine Hosey, Brian St Thomas, and Jean Garcia-Gathright. 2019. Search mindsets: Understanding focused and non-focused information seeking in music search. In *The World Wide Web Conference*. 2971–2977.

[102] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[103] Jianhua Lin. 2002. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory* 37, 1 (2002), 145–151.

[104] Siyi Lin, Chongming Gao, Jiawei Chen, Sheng Zhou, Binbin Hu, Yan Feng, Chun Chen, and Can Wang. 2025. How Do Recommendation Models Amplify Popularity Bias? An Analysis from the Spectral Perspective. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining* (Hannover, Germany) *(WSDM '25)*. Association for Computing Machinery, New York, NY, USA, 659–668. doi:10.1145/3701551.3703579

[105] Xinyu Lin, Yashar Deldjoo, Sunhao Dai, Honghui Bao, Xiaopeng Ye, Fatemeh Nazary, Wenjie Wang, Tommaso Di Noia, Jun Xu, and Tat-Seng Chua. 2026. Autonomous Information Seeking: A Roadmap for Agentic Recommender Systems. *arXiv preprint arXiv:2607.04433* (2026).

[106] Meijun Liu, Xiao Hu, and Markus Schedl. 2018. The relation of culture, socio-economics, and friendship to music preferences: A large-scale, cross-country study. *PLOS ONE* 13, 12 (Dec. 2018), e0208186. doi:10.1371/journal.pone.0208186 Publisher: Public Library of Science.

[107] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173.

[108] Ruilun Liu and Xiao Hu. 2020. A Multimodal Music Recommendation System with Listeners' Personality and Physiological Signals. In *JCDL '20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Virtual Event, China, August 1-5, 2020*, Ruhua Huang, Dan Wu, Gary Marchionini, Daqing He, Sally Jo Cunningham, and Preben Hansen (Eds.). ACM, 357–360. doi:10.1145/3383583.3398623

[109] Alex B Long. 2007. (Insert Song Lyrics Here): The Uses and Misuses of Popular Music Lyrics in Legal Writing. *Wash. & Lee L. Rev.* 64 (2007), 531.

[110] Do Xuan Long, Ngoc-Hai Nguyen, Tiviatis Sim, Hieu Dao, Shafiq Joty, Kenji Kawaguchi, Nancy F. Chen, and Min-Yen Kan. 2025. LLMs Are Biased Towards Output Formats! Systematically Evaluating and Mitigating Output Format Bias of LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 299–330. doi:10.18653/v1/2025.naacl-long.15

[111] Adam J. Lonsdale and Adrian C. North. 2011. Why do we listen to music? A uses and gratifications analysis. *British Journal of Psychology (London, England: 1953)* 102, 1 (Feb. 2011), 108–134. doi:10.1348/000712610X506831

[112] Feng Lu and Nava Tintarev. 2018. A Diversity Adjusting Strategy with Personality for Music Recommendation. In *Proceedings of the 5th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, IntRS 2018, co-located with ACM Conference on Recommender Systems (RecSys 2018), Vancouver, Canada, October 7, 2018 (CEUR Workshop Proceedings, Vol. 2225)*, Peter Brusilovsky, Marco de Gemmis, Alexander Felfernig, Pasquale Lops, John O'Donovan, Giovanni Semeraro, and Martijn C. Willemsen (Eds.). CEUR-WS.org, 7–14. https://ceur-ws.org/Vol-2225/paper2.pdf

[113] Malte Ludewig and Dietmar Jannach. 2018. Evaluation of session-based recommendation algorithms. *User Modeling and User-Adapted Interaction* 28, 4 (2018), 331–390.

[114] Kai Luo, Scott Sanner, Ga Wu, Hanze Li, and Hojin Yang. 2020. Latent linear critiquing for conversational recommender systems. In *Proceedings of The Web Conference 2020*. 2535–2541.

[115] Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, and Jiebo Luo. 2024. LLM-Rec: Personalized Recommendation via Prompting Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 583–612. doi:10.18653/v1/2024.findings-naacl.39

[116] Lucas S Maia, Martín Rocamora, Luiz WP Biscainho, and Magdalena Fuentes. 2022. Adapting meter tracking models to Latin American music. In *Ismir 2022 Hybrid Conference*.

[117] Reza Yousefi Maragheh and Yashar Deldjoo. 2026. The Future is Agentic: Definitions, Perspectives, and Open Challenges of Multi-Agent Recommender Systems. *ACM Transactions on Recommender Systems* (2026).

[118] Lilian Marey, Bruno Sguerra, and Manuel Moussallam. 2024. Modeling activity-driven music listening with pace. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*. 346–351.

[119] Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 4984–4997. doi:10.18653/v1/2020.acl-main.448

[120] Kristina Matrosova, Lilian Marey, Guillaume Salha-Galvan, Thomas Louail, Olivier Bodini, and Manuel Moussallam. 2024. Do recommender systems promote local music? a reproducibility study using music streaming data. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 148–157.

[121] Robert R. McCrae and Oliver P. John. 1992. An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality* 60, 2 (1992), 175–215. doi:10.1111/j.1467-6494.1992.tb00970.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-6494.1992.tb00970.x

[122] 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. In *Proceedings of the 19th ACM Conference on Recommender Systems (RecSys)*. ACM, Prague, Czhech Republic. Short Paper.

[123] Alessandro B. Melchiorre and Markus Schedl. 2020. Personality Correlates of Music Audio Preferences for Modelling Music Listeners. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 313–317. doi:10.1145/3340631.3394874

[124] MetaBrainz Foundation. [n. d.]. MusicBrainz Database. https://musicbrainz.org/doc/MusicBrainz_Database. Accessed 2025-10-11.

[125] MetaBrainz Foundation. n.d.. MusicBrainz: the open music encyclopedia. https://musicbrainz.org/. Accessed 2025-10-11.

[126] BSL METEOR. 2005. an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

[127] Erik Miehling, Michael Desmond, Karthikeyan Natesan Ramamurthy, Elizabeth M. Daly, Kush R. Varshney, Eitan Farchi, Pierre Dognin, Jesus Rios, Djallel Bouneffouf, Miao Liu, and Prasanna Sattigeri. 2025. Evaluating the Prompt Steerability of Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 7874–7900. doi:10.18653/v1/2025.naacl-long.400

[128] Sho Miyakawa and Takehito Utsuro. 2024. Emotion Classification of Lyrics through Summarization by Large Language Models. In *2024 IEEE International Conference on Big Data (BigData)*. 2999–3006. doi:10.1109/BigData62323.2024.10825406

[129] Lori A Morea. 2005. The Future of Music in a Digital Age: The Ongoing Conflict Between Copyright Law and Peer-to-Peer Technology. *Campbell L. Rev.* 28 (2005), 195.

[130] Meinard Müller. 2015. *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Vol. 5. Springer.

[131] Sheshera Mysore, Mahmood Jasim, Andrew McCallum, and Hamed Zamani. 2023. Editable user profiles for controllable text recommendations. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 993–1003.

[132] Sergio Oramas, Andres Ferraro, Alvaro Sarasua, and Fabien Gouyon. 2024. Talking to Your Recs: Multimodal Embeddings For Recommendation and Retrieval. In *Proceedings of the 2nd Music Recommender Systems Workshop 2024 co-located with the 18th ACM Conference on Recommender Systems (RecSys 2024)*.

[133] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[134] Emilia Parada-Cabaleiro, Anton Batliner, Marcel Zentner, and Markus Schedl. 2023. Exploring emotions in Bach chorales: a multi-modal perceptual and data-driven study. *Royal Society Open Science* 10, 12 (Dec. 2023), 230574. doi:10.1098/rsos.230574 Publisher: Royal Society.

[135] Donguk Park, Dongwon Lee, and Yeon-Chang Lee. 2026. Echoes in the Loop: Diagnosing Risks in LLM-Powered Recommender Systems under Feedback Loops. *arXiv preprint arXiv:2602.07442* (2026).

[136] Yoon-Joo Park and Alexander Tuzhilin. 2008. The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM conference on Recommender systems*. 11–18.

[137] Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 15012–15032.

[138] Geoffroy Peeters et al. 2004. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. *CUIDADO Ist Project Report* 54, 0 (2004), 1–25.

[139] Andreas Peintner, Marta Moscati, Yu Kinoshita, Richard Vogl, Peter Knees, Markus Schedl, Hannah Strauss, Marcel Zentner, and Eva Zangerle. 2025. Nuanced Music Emotion Recognition via a Semi-Supervised Multi-Relational Graph Neural Network. *Transactions of the International Society for Music Information Retrieval* 8, 1 (June 2025). doi:10.5334/tismir.235

[140] Maxime Peyrard. 2019. Studying Summarization Evaluation Metrics in the Appropriate Scoring Range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 5093–5100. doi:10.18653/v1/P19-1502

[141] Pantelis Pipergias Analytis and Philipp Hager. 2023. Collaborative filtering algorithms are prone to mainstream-taste bias. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 750–756.

[142] Liam Pond, Sichen Meng, Linnea Kirby, Simon Ngassam, Sebastien Chow, Dylan Hillerbrand, and Ichiro Fujinaga. 2025. SESEMMI for LinkedMusic: Democratizing Access to Musical Archives via Large Language Models. In *1st Workshop on Large Language Models for Music {\&} Audio (LLM4MA)*.

[143] Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. 2018. Sequence-aware recommender systems. *ACM computing surveys (CSUR)* 51, 4 (2018), 1–36.

[144] Filip Radlinski, Krisztian Balog, Fernando Diaz, Lucas Dixon, and Ben Wedin. 2022. On natural language user profiles for transparent and scrutable recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 2863–2874.

[145] Rashin Rahnamoun and Mehrnoush Shamsfard. 2025. Multi-Layered Evaluation Using a Fusion of Metrics and LLMs as Judges in Open-Domain Question Answering. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE,

6088–6104. https://aclanthology.org/2025.coling-main.408/

[146] Yves Raimond, Samer A. Abdallah, Mark B. Sandler, and Frederick Giasson. 2007. The Music Ontology. In *Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007, Vienna, Austria, September 23-27, 2007*, Simon Dixon, David Bainbridge, and Rainer Typke (Eds.). Austrian Computer Society, 417–422. http://ismir2007.ismir.net/proceedings/ISMIR2007_p417_raimond.pdf

[147] Rajeev Rajan, Joshua Antony, Riya Ann Joseph, Jijohn M. Thomas, Chandr Dhanush H, and Arunima C. V. 2021. Audio-Mood Classification Using Acoustic-Textual Feature Fusion. In *2021 Fourth International Conference on Microelectronics, Signals & Systems (ICMSS)*. 1–6. doi:10.1109/ICMSS53060.2021.9673592

[148] Jerome Ramos, Hossein A Rahmani, Xi Wang, Xiao Fu, and Aldo Lipani. 2024. Transparent and scrutable recommendations using natural language user profiles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 13971–13984.

[149] Brian Regan, Desislava Hristova, and Mariano Beguerisse-Díaz. 2023. Which Witch? Artist name disambiguation and catalog curation using audio and metadata. https://research.atspotify.com/2023/11/which-witch-artist-name-disambiguation-and-catalog-curation-using-audio-and-metadata. Published by Spotify Research.

[150] Lise Regnier and Geoffroy Peeters. 2009. Singing voice detection in music tracks using direct voice vibrato detection. In *2009 IEEE international conference on acoustics, speech and signal processing*. IEEE, 1685–1688.

[151] Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are References Really Needed? Unbabel-IST 2021 Submission for the Metrics Shared Task. In *Proceedings of the Sixth Conference on Machine Translation*, Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz (Eds.). Association for Computational Linguistics, Online, 1030–1040. https://aclanthology.org/2021.wmt-1.111/

[152] Peter J Rentfrow. 2012. The role of music in everyday life: Current directions in the social psychology of music. *Social and personality psychology compass* 6, 5 (2012), 402–416.

[153] Peter J. Rentfrow and Samuel D. Gosling. 2007. The content and validity of music-genre stereotypes among college students. *Psychology of Music* 35, 2 (2007), 306–326. doi:10.1177/0305735607070382

[154] Renata L Rosa, Demsteneso Z Rodriguez, and Graça Bressan. 2015. Music recommendation system based on user's sentiments extracted from social networks. *IEEE Transactions on Consumer Electronics* 61, 3 (2015), 359–367.

[155] Jia-Jia Ruan, Xi-Xu He, Min Zhang, and Yuan Gao. 2023. Entity generation algorithm based on reference expansion. *Journal of Electronic Science and Technology* 21, 3 (2023), 100218.

[156] Milad Sabouri, Masoud Mansoury, Kun Lin, and Bamshad Mobasher. 2025. Towards Explainable Temporal User Profiling with LLMs. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*. 219–227.

[157] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

[158] Ananya B. Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. 2021. Perturbation CheckLists for Evaluating NLG Evaluation Metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 7219–7234. doi:10.18653/v1/2021.emnlp-main.575

[159] Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. GoLLIE: Annotation Guidelines improve Zero-Shot Information-Extraction. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=Y3wpuxd7u9

[160] Krishna Sayana, Raghavendra Vasudeva, Yuri Vasilevski, Kun Su, Liam Hebert, James Pine, Hubert Pham, Ambarish Jash, and Sukhdeep Sodhi. 2025. Beyond Retrieval: Generating Narratives in Conversational Recommender Systems. In *Companion Proceedings of the ACM on Web Conference 2025* (Sydney NSW, Australia) *(WWW '25)*. Association for Computing Machinery, New York, NY, USA, 2411–2420. doi:10.1145/3701716.3717531

[161] Thomas Schäfer. 2016. The goals and effects of music listening and their relationship to the strength of music preference. *PloS one* 11, 3 (2016), e0151634.

[162] Markus Schedl, Emilia Gómez, Julián Urbano, et al. 2014. Music information retrieval: Recent developments and applications. *Foundations and Trends® in Information Retrieval* 8, 2-3 (2014), 127–261.

[163] Markus Schedl, Emilia Gómez, Erika S. Trent, Marko Tkalčič, Hamid Eghbal-Zadeh, and Agustín Martorell. 2018. On the Interrelation Between Listener Characteristics and the Perception of Emotions in Classical Orchestra Music. *IEEE Transactions on Affective Computing* 9, 4 (Oct. 2018), 507–525. doi:10.1109/TAFFC.2017.2663421

[164] Markus Schedl, Peter Knees, Brian McFee, and Dmitry Bogdanov. 2021. *Music Recommendation Systems: Techniques, Use Cases, and Challenges*. Springer US, 927–971. doi:10.1007/978-1-0716-2197-4_24

[165] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. 2018. Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval* 7, 2 (2018), 95–116.

[166] Thomas Schäfer, Peter Sedlmeier, Christine Städtler, and David Huron. 2013. The psychological functions of music listening. *Frontiers in Psychology* 4 (Aug. 2013). doi:10.3389/fpsyg.2013.00511 Publisher: Frontiers.

[167] Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization Asks for Fact-based Evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6594–6604. doi:10.18653/v1/2021.emnlp-main.529

[168] Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 7881–7892. doi:10.18653/v1/2020.acl-main.704

[169] Bruno Sguerra, Marion Baranes, Romain Hennequin, and Manuel Moussallam. 2022. Navigational, informational or punk-rock? An exploration of search intent in the musical domain. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. 202–211.

[170] Bruno Sguerra, Elena V Epure, Harin Lee, and Manuel Moussallam. 2025. Biases in LLM-Generated Musical Taste Profiles for Recommendation. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*. 527–532.

[171] Bruno Sguerra, Viet-Anh Tran, and Romain Hennequin. 2022. Discovery dynamics: Leveraging repeated exposure for user and music characterization. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 556–561.

[172] Bruno Sguerra, Viet-Anh Tran, and Romain Hennequin. 2023. Ex2Vec: Characterizing users and items from the mere exposure effect. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 971–977.

[173] Bruno Sguerra, Viet-Anh Tran, Romain Hennequin, and Manuel Moussallam. 2025. Uncertainty in Repeated Implicit Feedback as a Measure of Reliability. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*. 234–242.

[174] Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. 2022. On the Evaluation Metrics for Paraphrase Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3178–3190. doi:10.18653/v1/2022.emnlp-main.208

[175] Tiancheng Shen, Jia Jia, Yan Li, Yihui Ma, Yaohua Bu, Hanjie Wang, Bo Chen, Tat-Seng Chua, and Wendy Hall. 2020. PEIA: Personality and Emotion Integrated Attentive Model for Music Recommendation on Social Media Platforms. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 206–213. doi:10.1609/AAAI.V34I01.5352

[176] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2219–2228.

[177] Yading Song, Simon Dixon, Marcus T Pearce, and Andrea R Halpern. 2016. Perceived and induced emotion responses to popular music: Categorical and dimensional models. *Music Perception: An Interdisciplinary Journal* 33, 4 (2016), 472–492.

[178] Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. 2025. The Good, The Bad, and The Greedy: Evaluation of LLMs Should Not Ignore Non-Determinism. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 4195–4206. doi:10.18653/v1/2025.naacl-long.211

[179] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM conference on recommender systems*. 154–162.

[180] Lei Sun, Jinming Zhao, and Qin Jin. 2024. Revealing Personality Traits: A New Benchmark Dataset for Explainable Personality Recognition on Dialogues. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 19988–20002. doi:10.18653/v1/2024.emnlp-main.1115

[181] Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. 2022. BERTScore is Unfair: On Social Bias in Language Model-Based Metrics for Text Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3726–3739. doi:10.18653/v1/2022.emnlp-main.245

[182] Noah Tekle, Alline Ayala, Jonathan Haile, Abdulla Alshabanah, Corey Baker, and Murali Annavaram. 2024. Music Recommendation through LLM Song Summary. *ACM The 1st Workshop on Risks, Opportunities, and Evaluation of Generative Models in Recommender Systems (ROEGEN@RecSys2024)* 114 (2024), 71–690.

[183] Brian Thompson and Matt Post. 2020. Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 90–121. doi:10.18653/v1/2020.emnlp-main.8

[184] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems: Methodological issues and empirical studies on the impact of personalization. *User Modeling and User-Adapted Interaction* 22, 4 (2012), 399–439.

[185] Nava Tintarev and Judith Masthoff. 2015. Explaining recommendations: Design and evaluation. In *Recommender systems handbook*. Springer, 353–382.

[186] Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: an analysis of language models on negation benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, Alexis Palmer and Jose Camacho-collados (Eds.). Association for Computational Linguistics, Toronto, Canada, 101–114. doi:10.18653/v1/2023.starsem-1.10

[187] Alicia Tsai, Adam Kraft, Long Jin, Chenwei Cai, Anahita Hosseini, Taibai Xu, Zemin Zhang, Lichan Hong, Ed Chi, and Xinyang Yi. 2024. Leveraging LLM Reasoning Enhances Personalized Recommender Systems. In *Findings of the Association for Computational Linguistics ACL 2024*. 13176–13188.

[188] Robin Ungruh, Karlijn Dinnissen, Anja Volk, Maria Soledad Pera, and Hanna Hauptmann. 2024. Putting Popularity Bias Mitigation to the Test: A User-Centric Evaluation in Music Recommenders. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 169–178.

[189] Saúl Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*. ACM, Chicago, IL, USA, 109–116. doi:10.1145/2043932.2043955

[190] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.

[191] et al. Wang. 2025. Diagnostic-Guided Dynamic Profile Optimization for LLM-based User Simulation. *arXiv preprint arXiv:2508.12645* (2025).

[192] Jiarong Wang, Jiaji Wu, Mingzhou Tan, and Lingxuan Zhu. 2025. Emotion-Aware Conversational Music Recommendation With Multiagent System. *IEEE Transactions on Computational Social Systems* (2025), 1–14. doi:10.1109/TCSS.2025.3599008

[193] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 9440–9450. doi:10.18653/v1/2024.acl-long.511

[194] Shijia Wang, Tianpei Ouyang, Yunfan Zhou, Qiang Xiao, Yintao Ren, Yifei Pan, Fangjian Li, and Chuanjiang Luo. 2025. Enhanced Emotion-aware Music Recommendation via Large Language Models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (Toronto ON, Canada) *(KDD '25)*. Association for Computing Machinery, New York, NY, USA, 4986–4994. doi:10.1145/3711896.3737212

[195] Xuezhi Wang, Jason Wei, Dale Schuurmans, Tobias Bosma, and Quoc V. Le. 2023. Self-Consistency Improves Chain of Thought Reasoning in Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*. https://arxiv.org/abs/2203.11171 arXiv preprint arXiv:2203.11171.

[196] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods* 45, 4 (2013), 1191–1207.

[197] William Webber, Alistair Moffat, and Justin Zobel. 2010. A Similarity Measure for Indefinite Rankings. In *ACM Transactions on Information Systems*, Vol. 28. ACM, 1–38. doi:10.1145/1852102.1852106

[198] Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, György Fazekas, and Dmitry Bogdanov. 2024. MuChoMusic: Evaluating Music Understanding in Multimodal Audio-Language Models. In *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*.

[199] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[200] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. 2024. A survey on large language models for recommendation. *World Wide Web* 27, 5 (2024), 60.

[201] Shangda Wu, Guo Zhancheng, Ruibin Yuan, Junyan Jiang, Seungheon Doh, Gus Xia, Juhan Nam, Xiaobing Li, Feng Yu, and Maosong Sun. 2025. Clamp 3: Universal music information retrieval across unaligned modalities and unseen languages. In *Findings of the Association for Computational Linguistics: ACL 2025*. 2605–2625.

[202] Meng Yang, Jon McCormack, Maria Teresa Llano, and Wanchao Su. 2025. Exploring the Feasibility of LLMs for Automated Music Emotion Annotation. In *Ismir 2025 Hybrid Conference*.

[203] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems* 36 (2023), 11809–11822.

[204] Chao Yu, Qixin Tan, Hong Lu, Jiaxuan Gao, Xinting Yang, Yu Wang, Yi Wu, and Eugene Vinitsky. 2025. ICPL: Few-shot In-context Preference Learning via LLMs. arXiv:2410.17233 [cs.AI] https://arxiv.org/abs/2410.17233

[205] Yakun Yu, Shi-ang Qi, Baochun Li, and Di Niu. 2024. PepRec: Progressive Enhancement of Prompting for Recommendation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 17941–17953. doi:10.18653/v1/2024.emnlp-main.995

[206] Ruibin Yuan, Hanfeng Lin, Yi Wang, Zeyue Tian, Shangda Wu, Tianhao Shen, Ge Zhang, Yuhang Wu, Cong Liu, Ziya Zhou, Liumeng Xue, Ziyang Ma, Qin Liu, Tianyu Zheng, Yizhi Li, Yinghao Ma, Yiming Liang, Xiaowei Chi, Ruibo Liu, Zili Wang, Chenghua Lin, Qifeng Liu, Tao Jiang, Wenhao Huang, Wenhu Chen, Jie Fu, Emmanouil Benetos, Gus Xia, Roger Dannenberg, Wei Xue, Shiyin Kang, and Yike Guo. 2024. ChatMusician: Understanding and Generating Music Intrinsically with LLM. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 6252–6271. doi:10.18653/v1/2024.findings-acl.373

[207] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTSCORE: evaluating generated text as text generation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*. Curran Associates Inc., Red Hook, NY, USA, Article 2088, 15 pages.

[208] Sojeong Yun and Youn-kyung Lim. 2025. User Experience with LLM-powered Conversational Recommendation Systems: A Case of Music Recommendation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 898, 15 pages. doi:10.1145/3706598.3713347

[209] Eva Zangerle and Christine Bauer. 2022. Evaluating recommender systems: survey and framework. *ACM computing surveys* 55, 8 (2022), 1–38.

[210] Eva Zangerle, Martin Pichl, and Markus Schedl. 2020. User Models for Culture-Aware Music Recommendation: Fusing Acoustic and Cultural Cues. *Transactions of the International Society for Music Information Retrieval* 3, 1 (March 2020). doi:10.5334/tismir.37

[211] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1569–1578.

[212] Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating Factual Consistency with A Unified Alignment Function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 11328–11348. doi:10.18653/v1/2023.acl-long.634

[213] Baiqiao Zhang, Zhifeng Liao, Xiangxian Li, Chao Zhou, Juan Liu, Xiaojuan Ma, and Yulong Bian. 2025. Rethinking Personality Assessment from Human-Agent Dialogues: Fewer Rounds May Be Better Than More. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 5357–5380. https://aclanthology.org/2025.findings-emnlp.287/

[214] Hanlin Zhang, YiFan Zhang, Yaodong Yu, Dhruv Madeka, Dean Foster, Eric Xing, Himabindu Lakkaraju, and Sham Kakade. 2024. A Study on the Calibration of In-context Learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 6118–6136. doi:10.18653/v1/2024.naacl-long.340

[215] Jie Zhang, Yanan Lu, Katayoun Mehdinezhadnouri, Junjie Liu, and Haohui Lu. 2025. Impact of music-based interventions on subjective well-being: a meta-analysis of listening, training, and therapy in clinical and nonclinical populations. *Frontiers in Psychology* 16 (2025), 1608508.

[216] Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. LLMaAA: Making Large Language Models as Active Annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 13088–13103. doi:10.18653/v1/2023.findings-emnlp.872

[217] Tong Zhang. 2025. AdaptRec: A Self-Adaptive Framework for Sequential Recommendations with Large Language Models. arXiv:2504.08786 [cs.IR] https://arxiv.org/abs/2504.08786

[218] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr

[219] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 563–578.

[220] Wei Zhao, Michael Strube, and Steffen Eger. 2023. DiscoScore: Evaluating Text Generation with BERT and Discourse Coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 3865–3883. doi:10.18653/v1/2023.eacl-main.278

[221] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, Vol. 36. 46595–46623.

[222] Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine A Heller, and Subhrajit Roy. 2024. Batch Calibration: Rethinking Calibration for In-Context Learning and Prompt Engineering. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=L3FHMoKZcS

[223] Joyce Zhou, Yijia Dai, and Thorsten Joachims. 2024. Language-Based User Profiles for Recommendation. *arXiv preprint arXiv:2402.15623* (2024).

[224] Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 1950–1966.

[225] Le Zhuo, Ruibin Yuan, Jiahao Pan, Yinghao Ma, Yizhi Li, Ge Zhang, Si Liu, Roger Dannenberg, Jie Fu, Chenghua Lin, et al. 2023. Lyricwhiz: Robust multilingual zero-shot lyrics transcription by whispering to chatgpt. *arXiv preprint arXiv:2306.17103* (2023).

[226] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*. 22–32.