

# Music Playlist Captioning at Scale with Large Language Models

Mathieu Delcluze<sup>1</sup>, Léa Briand<sup>1</sup>, Benjamin Chapus<sup>1</sup>,  
Deniz Mekik<sup>1</sup>, and Guillaume Salha-Galvan<sup>2</sup>

<sup>1</sup> Deezer Research

<sup>2</sup> SJTU Paris Elite Institute of Technology  
`research@deezer.com`

**Abstract.** Music streaming services such as Deezer often recommend personalized playlists to users. Playlist captioning, which involves describing these playlists in natural language, is essential for helping users understand the content behind each recommendation, yet remains challenging at scale. This paper presents the automatic playlist captioning system deployed on Deezer in 2025 to address this challenge. Leveraging recent advances in large language models (LLMs) to generate descriptive captions from diverse data sources in a controlled manner, this system now powers the Daily Mix feature, used by millions of users. This deployment has led to significant improvements in user engagement, highlighting how the semantic framing of an unchanged recommendation shapes user perception in online personalized experiences.

**Keywords:** Music Playlist Captioning · Large Language Model · Music Streaming Service · Recommender System · A/B Testing.

## 1 Introduction

Deezer is a global music streaming service, offering access to over 120 million tracks to 18 million active users across 180 countries [14]. To help users navigate this vast music catalog, the service relies on large-scale recommender systems, which are essential for driving user engagement and retention [2, 4, 6–8]. In particular, Deezer’s homepage features *Daily Mix* playlists, a daily set of personalized music playlists based on each user’s listening preferences.

A key challenge with Daily Mix is ensuring that users understand the content behind each recommended playlist. As shown in Figure 1, Deezer represents these collections using images combining two album covers along with a short list of selected featured artists. Adding descriptive titles could provide meaningful context. However, summarizing playlists in natural language remains a challenging task, known as playlist captioning [12]. This difficulty is amplified in industry settings, characterized by a large number of playlists.

This paper presents the automatic playlist captioning system deployed on Deezer in 2025 to address this challenge. Leveraging recent advances in large

language models (LLMs) [18, 33], this production-scale solution generates descriptive playlist captions from diverse data sources in a controlled yet flexible manner. A large-scale online A/B test on millions of Daily Mix users shows significant improvements in user engagement with recommended playlists. Beyond performance gains, our results underscore how semantic framing plays a critical role in shaping user perception of recommendations, even when the underlying content remains the same.

The remainder of this paper is organized as follows. Section 2 presents the Daily Mix feature in more detail, along with its captioning challenges. Section 3 introduces our method, including the motivation for using LLMs, the main components of our pipeline, and the validation and safeguard procedures adopted before deployment. Section 4 reports and discussed our online A/B test results on Deezer, and Section 5 concludes.

## 2 Recommending Daily Mix Playlists on Deezer

We begin this section by providing a broader overview of the Daily Mix feature, before discussing the challenges of presenting playlist content to users.

### 2.1 The Daily Mix Feature

The rise of music streaming services has led to the emergence of playlists as a primary way to listen to music [4, 23, 31]. To help users discover and enjoy content through this format, Deezer offers the Daily Mix feature. This recommender system, prominently displayed on the Deezer homepage, presents five personalized playlists daily to each user, each containing fifty music tracks. These playlists are represented by illustrative covers in a swipeable carousel [3], allowing users to easily browse and click to launch them.

Daily Mix playlists are generated based on each user’s listening preferences. More precisely, a proprietary algorithm<sup>3</sup> combines collaborative filtering and content-based techniques [25, 31] to construct a large graph of similar artists. A Louvain-inspired scalable community detection algorithm [5] extracts hierarchical clusters of similar artists from this graph. Daily Mix then associates users with clusters, considering both (1) relevance, based on user-cluster affinities according to a combination of usage signals, and (2) cluster diversity, using the Maximum Marginal Relevance (MMR) method [10, 32]. Finally, Daily Mix generates playlists from each cluster by selecting tracks from artists within the cluster, aiming to balance tracks already familiar to the user with new discoveries.

### 2.2 Challenges in Daily Mix Playlist Captioning

Regardless of recommendation quality, a key challenge with Daily Mix is ensuring that users understand the content of each playlist. As shown in Figure 1,

<sup>3</sup> Some technical details are omitted in this paper for confidentiality reasons.

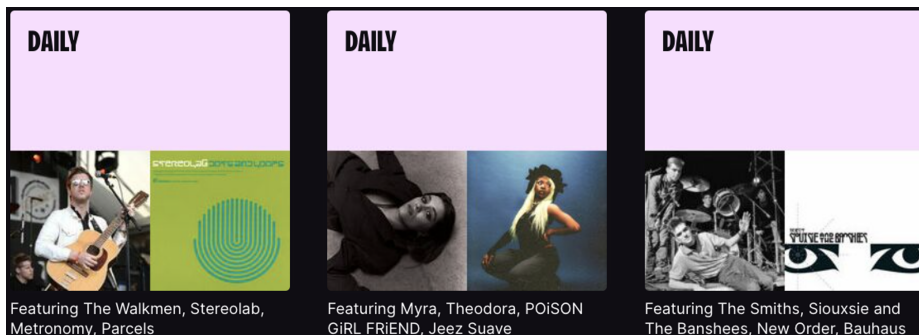


Fig. 1. Prior presentation of Daily Mix playlists on Deezer until 2025.

playlist illustrations combine two album covers with a short list of featured artists, providing only a rough and often incomplete view of each playlist’s identity. Moreover, the generic "Daily" titles are uninformative.

Replacing these titles with descriptive captions, similar to those used for Deezer’s editorial playlists curated by professionals (e.g., "Modern Rock Essentials"), could clarify each playlist’s content. However, given the scale and dynamic nature of Daily Mix, providing human-written captions is infeasible. In addition, automatically generating captions remains a technically challenging task, known as playlist captioning. Introduced by Choi et al. [12], it has since been approached using various deep neural network architectures that aim to model signals such as audio features, linguistic and musicological knowledge, and artist-level side information [12, 17, 20, 21, 24].

Nonetheless, prior work still highlights the difficulty of this task. Training data, particularly user-created playlists, is often noisy and lacks consistent templates. A semantic gap also persists when playlist concepts cannot be inferred from track embeddings alone, while linguistic fluency and the inherent subjectivity of captions add further complexity [21, 24]. Finally, although streaming services often describe music using labels (e.g., genres), which can support captioning, these descriptions remain subjective, incomplete, and sometimes inaccurate [11, 19, 26]. Together, these factors make playlist captioning particularly challenging in large-scale industrial settings.

### 3 LLM-Based Daily Mix Playlist Captioning

In this section, we present our LLM-based method for large-scale automatic playlist captioning in Daily Mix. We first motivate the use of LLMs for this task, to our knowledge for the first time, before presenting our technical pipeline. We then describe a set of caption validation and safety checks adopted internally prior to deployment.

### 3.1 Motivations for LLM-Based Captioning

Unlike previous work, we investigate playlist captioning using LLMs, which are powerful neural models trained on large corpora and capable of generating coherent text from diverse inputs [1, 13]. This decision is driven by growing evidence of their potential to enhance recommender systems, including by serving as content annotators and summarizers [18, 28, 33–35]. Recent work, in particular, shows that LLMs outperform baselines in the related task of music track captioning [9, 15, 16, 29]. We refer to Epure et al. [18] for a review of LLM applications and challenges in music recommendation specifically.

Our motivation is reinforced by an internal analytics study of user-created Deezer playlists, which revealed significant variability in titles. The tens of thousands of playlists created daily cover a wide range of themes, from genres and artists to periods, occasions, moods, activities, personal events, and ambiguous concepts, with titles varying in length and structure. This diversity underscores the need for a flexible captioning solution, making LLMs ideal candidates.

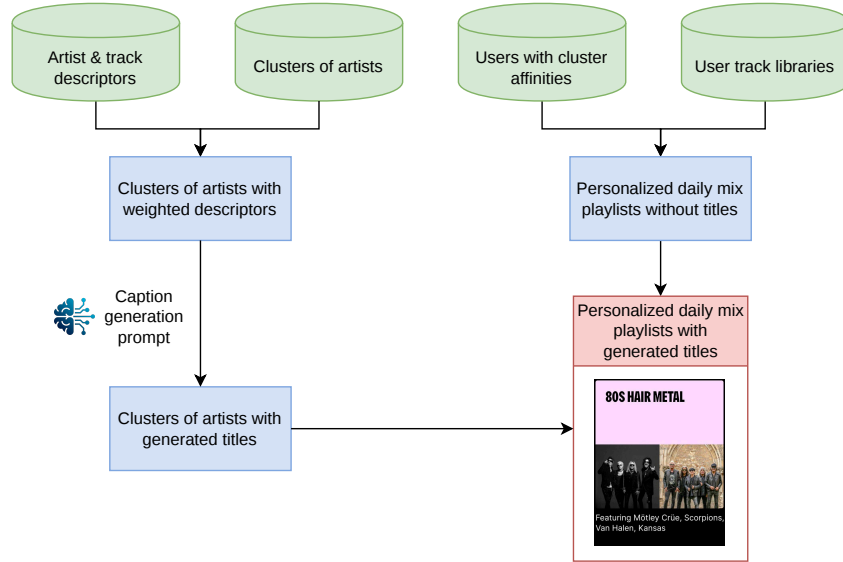
### 3.2 Pipeline for LLM-Based Captioning

**Artist Cluster Tagging for LLM Input** We now present our technical pipeline, illustrated in Figure 2. As explained in Section 2.1, Daily Mix playlists are built from clusters of similar artists. To maintain scalability in playlist captioning, we operate one step upstream of playlist generation and choose to generate captions for the approximately 5,000 existing artist clusters on Deezer, rather than for the millions of personalized playlists produced daily. This approach significantly reduces the computational load in our production environment.

For each cluster, we first associate a range of labels, or *tags*, from Deezer’s proprietary catalog, linking these tags to tracks and artists. These tags cover various descriptive aspects of music, including genres [19, 22], moods [6], countries [30], and decades. They are obtained through a combination of complementary methods, including audio signal analysis, textual analysis of artist biographies and other music descriptions, and manual annotations by professional editors. Each tag-track pair is assigned a confidence weight.

In addition, artists from each cluster appear in user-created playlists on Deezer. We associate information from titles of these playlists with each cluster, as user-written titles tend to be more creative and original than standard tags, providing a complementary perspective. To this end, we clean the titles by normalizing the text, removing stop-words, and filtering out short or non-informative titles, e.g., those containing only artist names. After cleaning, we compute the frequency and weight of each descriptor associated with tracks and aggregate them at the artist cluster level.

**Prompt Design and LLM Captioning** The steps outlined above associate, for each cluster, a ranked list of weighted descriptors, which we combine with a list of representative artists. This data serves as input for a carefully designed prompt, which instructs an LLM to act as a creative playlist curator. This LLM



**Fig. 2.** Pipeline overview of music playlist captioning for personalized Daily Mix recommendations in our production environment at Deezer.

initiates a multi-step process, where it first analyzes the input lists, identifies descriptive elements, and cross-checks them with artist information.

The LLM then generates short and evocative playlist titles in five key languages for Deezer (English, French, German, Portuguese, and Spanish), adhering to formatting constraints specified by product design. The prompt ensures that titles focus on specific genres, regions, and thematic elements while filtering out irrelevant or conflicting signals. It also encourages the detection of unifying themes, such as festivals or events, that cohesively link the tracks. Strict rules ensure consistency across languages and maintain a concise format, with titles limited to a maximum of 16 characters.

**Deployment** The pipeline is designed for seamless integration into large-scale production environments. In practice, we use Spark/Scala for data preparation and Python for artist clustering and playlist captioning. The LLM we employ is Gemini 2.0 Flash [13], chosen primarily for its cost-effectiveness, based on internal benchmarking of various models. Daily Mix playlists and their captions are generated offline daily and exported to a Cassandra cluster, from which they can be displayed at scale on Deezer.

### 3.3 Captioning Validation and Safeguards

Before being exposed to users, playlist captions had to undergo a series of internal quality and safeguard validation checks. First and foremost, our LLM prompt explicitly prevents the generation of offensive or inappropriate content. An extensive human study involving professional editors from Deezer also evaluated the compliance and faithfulness of captions before exposure to users.

We complemented the human validation with an LLM-as-a-Judge [27] verification procedure. Another instance of our LLM ran a second prompt, instructing it to act as a music taxonomy and catalog quality expert. Given playlist captions and tags, the LLM assigned quality scores between 0 and 10, reflecting how accurately each caption described the playlist, with an average score of 7.7/10.

A consequence of these checks was that we deployed the model for only three of the five languages (English, French, Portuguese), as the other two did not meet strict quality standards yet. This quality difference may be explained by the overrepresentation of English, French, and Portuguese speakers on Deezer, leading to more user playlists in these languages and thus more annotations.

## 4 Experimental Evaluation

We now report and discuss our experimental evaluation of the method on Deezer.

### 4.1 Experimental Setting

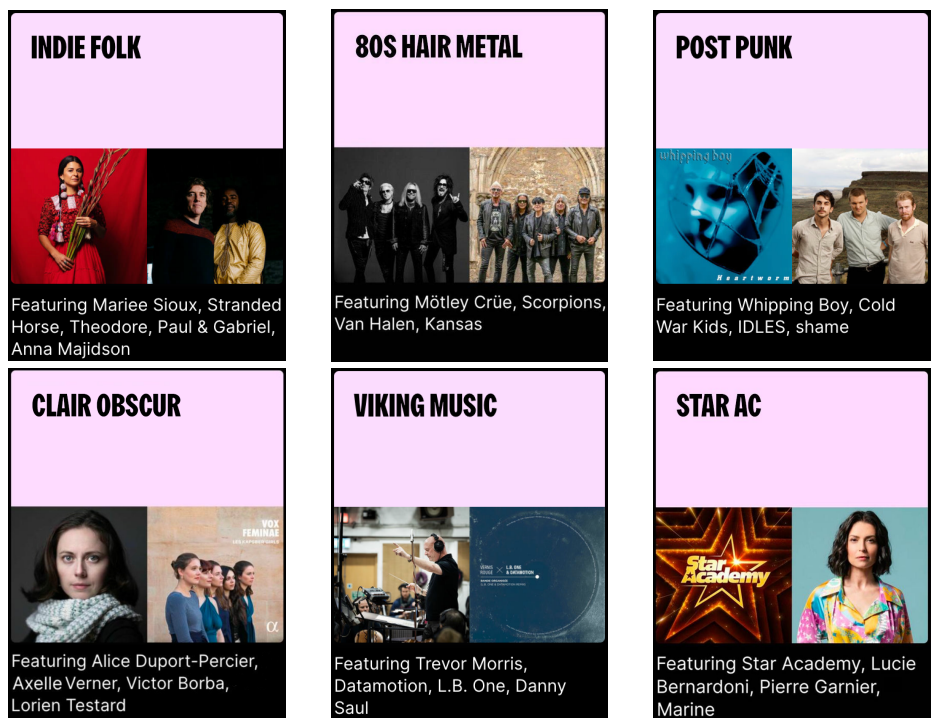
In September and October 2025, we conducted an online A/B test on millions of Deezer users. During this test, a randomly selected cohort received LLM-generated Daily Mix captions, while a control group experienced the original system. We compared both approaches using three performance indicators for user engagement: *adoption*, defined as the proportion of users who streamed a recommended playlist at least once during the test; *reconnection*, the percentage of users who streamed in at least three different weeks; and *satisfaction*, the percentage of users who added tracks to their favorites or a playlist, or exhibited long listening times with low skip rates, according to internal thresholds.

### 4.2 Results and Discussion

Table 1 shows the relative performance of LLM-based captioning compared to the baseline. Adding descriptive captions led to significant improvements in all metrics: a 24.9% increase in adoption, a 16.9% increase in reconnection, and an 11.5% increase in satisfaction (p-value < 0.01 for all metrics). Importantly, we recall that the Daily Mix algorithm was the same for both cohorts; only the captions differed. This highlights how semantic framing can strongly influence user behavior and perception of the same recommendations. In this case, adding descriptions helped users better understand each playlist and drove double-digit engagement gains without modifying the recommender system.

**Table 1.** Online A/B test on the Daily Mix feature: Relative improvements of LLM-generated playlist captions vs. baseline (p-value < 0.01 for all metrics).

Daily Mix with LLM captioning	Online A/B Test Metric		
	Adoption	Reconnection	Satisfaction
	+24.9%	+16.9%	+11.5%



**Fig. 3.** Examples of LLM-generated Daily Mix captions on Deezer.

Our LLM-as-a-Judge reveals that, in practice, users are primarily exposed to playlists rated over 8/10, with most scoring 9/10 or 10/10. Popular content on Deezer tends to be associated with richer descriptions during the annotation phase of Section 3.2, resulting in well-graded captions for playlists including this content. Additionally, Figure 3 illustrates that, while the LLM produces genre-based captions (e.g., *Indie Folk*), it also demonstrates flexibility and contextual relevance that goes beyond simple genre-based tagging. For example, *Clair Obscur* accurately describes a playlist of tracks from the video game "Clair Obscur: Expedition 33", while *Star Ac* captions a playlist of artists from a French TV show of the same name.

In summary, our test highlights the impact of playlist captioning on user perception and engagement, while confirming the utility of our LLM. Its flexibility is a clear strength. However, although no hallucinations were observed, we em-

phasize that this risk requires careful monitoring. In future work, we will aim to further assess online performance against non-LLM-based methods, benchmark additional LLMs beyond Gemini 2.0 Flash, and improve artist cluster tagging, a key factor for high-quality captions, particularly for less popular content and underrepresented languages.

## 5 Conclusion

In this work, we presented a production-scale automatic playlist captioning system leveraging recent advances in LLMs to generate descriptive captions from diverse data sources in a controlled manner. Through large-scale online A/B testing, we demonstrated significant improvements in user engagement with Daily Mix playlists. These results highlighted the critical role of semantic framing in shaping user perception of recommendations, even when the underlying content remains the same. Our findings also emphasized the flexibility of LLM-based captioning in capturing a wide range of musical themes and contexts. At the same time, our work emphasized the importance of robust validation procedures in real-world deployments, as well as the inherent risk of hallucinations. Based on these positive results, this system has been fully deployed on Deezer since late 2025. It now serves millions of users daily. Looking forward, we outlined several directions to further strengthen our approach. These directions open avenues for future research to further assess the potential of LLM-based captioning.

## References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bendada, W., Bontempelli, T., Morlon, M., Chapus, B., Cador, T., Bouabça, T., Salha-Galvan, G.: Track Mix Generation on Music Streaming Services using Transformers. In: Proceedings of the 17th ACM Conference on Recommender Systems. pp. 112–115 (2023)
3. Bendada, W., Salha, G., Bontempelli, T.: Carousel personalization in music streaming apps with contextual bandits. In: Proceedings of the 14th ACM Conference on Recommender Systems. pp. 420–425 (2020)
4. Bendada, W., Salha-Galvan, G., Bouabça, T., Cazenave, T.: A scalable framework for automatic playlist continuation on music streaming services. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 464–474 (2023)
5. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**(10), P10008 (2008)
6. Bontempelli, T., Chapus, B., Rigaud, F., Morlon, M., Lorant, M., Salha-Galvan, G.: Flow Moods: Recommending Music by Moods on Deezer. In: Proceedings of the 16th ACM Conference on Recommender Systems. pp. 452–455 (2022)

7. Briand, L., Bontempelli, T., Bendada, W., Morlon, M., Rigaud, F., Chapus, B., Bouabça, T., Salha-Galvan, G.: Let's get it started: Fostering the discoverability of new releases on deezer. In: Proceedings of the 46th European Conference on Information Retrieval. pp. 286–291. Springer (2024)
8. Briand, L., Salha-Galvan, G., Bendada, W., Morlon, M., Tran, V.A.: A Semi-Personalized System for User Cold Start Recommendation on Music Streaming Apps. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 2601–2609 (2021)
9. Bukey, I., Wang, Z., Donahue, C., Bryan, N.J.: Rethinking music captioning with music metadata llms. arXiv preprint arXiv:2602.03023 (2026)
10. Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 335–336 (1998)
11. Choi, J., Khlif, A., Epure, E.: Prediction of user listening contexts for music playlists. In: Proceedings of the 1st Workshop on NLP for Music and Audio. pp. 23–27 (2020)
12. Choi, K., Fazekas, G., McFee, B., Cho, K., Sandler, M.: Towards music captioning: Generating music playlist descriptions. In: Late-Breaking/Demo session of 17th International Society of Music Information Retrieval Conference (2016)
13. Comanici, G., Bieber, E., Schaeckermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
14. Deezer: <https://www.deezer.com> (2026)
15. Deng, Z., Ma, Y., Liu, Y., Guo, R., Zhang, G., Chen, W., Huang, W., Benetos, E.: 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. pp. 3643–3655 (2024)
16. Doh, S., Choi, K., Lee, J., Nam, J.: Lp-musiccaps: Llm-based pseudo music captioning. In: Proceedings of the 24th International Society for Music Information Retrieval Conference. vol. 2023, pp. 409–416 (2023)
17. Doh, S., Lee, J., Nam, J.: Music playlist title generation: A machine-translation approach. In: Proceedings of the 2nd Workshop on NLP for Music and Spoken Audio. pp. 27–31 (2021)
18. Epure, E.V., Deldjoo, Y., Sguerra, B., Schedl, M., Moussallam, M.: Music recommendation with large language models: Challenges, opportunities, and evaluation. arXiv preprint arXiv:2511.16478 (2025)
19. Epure, E.V., Salha, G., Hennequin, R.: Multilingual music genre embeddings for effective cross-lingual music item annotation. In: Proceedings of the 31st International Society for Music Information Retrieval Conference (2020)
20. Gabbolini, G., Bridge, D.: Surveying more than two decades of music information retrieval research on playlists. *ACM Transactions on Intelligent Systems and Technology* **15**(6), 1–68 (2024)
21. Gabbolini, G., Hennequin, R., Epure, E.: Data-efficient playlist captioning with musical and linguistic knowledge. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 11401–11415 (2022)
22. Hennequin, R., Royo-letelier, J., Moussallam, M.: Audio Based Disambiguation of Music Genre Tags. In: Proceedings of the 18th Conference of the International Society of Music Information Retrieval. pp. 645–653 (2018)

23. Jakobsen, L.: Playlists overtake albums in listenership, says loop study. Blog post on the Music Business Association (2016)
24. Kim, H., Doh, S., Lee, J., Nam, J.: Music playlist title generation using artist information. In: AAAI 2023 Workshop on Creative AI Across Modalities (2023)
25. Koren, Y., Rendle, S., Bell, R.: Advances in collaborative filtering. *Recommender Systems Handbook* pp. 91–142 (2021)
26. Lamere, P.: Social tagging and music information retrieval. *Journal of New Music Research* **37**(2), 101–114 (2008)
27. Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T., et al.: From generation to judgment: Opportunities and challenges of llm-as-a-judge. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. pp. 2757–2791 (2025)
28. Lin, J., Dai, X., Xi, Y., Liu, W., Chen, B., Zhang, H., Liu, Y., Wu, C., Li, X., Zhu, C., et al.: How can recommender systems benefit from large language models: A survey. *ACM Transactions on Information Systems* **43**(2), 1–47 (2025)
29. Liu, S., Hussain, A.S., Sun, C., Shan, Y.: Music understanding llama: Advancing text-to-music generation with question answering and captioning. In: *Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*. pp. 286–290. IEEE (2024)
30. Salha-Galvan, G., Hennequin, R., Chapus, B., Tran, V.A., Vazirgiannis, M.: Cold start similar artists ranking with gravity-inspired graph autoencoders. In: *Proceedings of the 15th ACM Conference on Recommender Systems*. pp. 443–452 (2021)
31. Schedl, M., Zamani, H., Chen, C.W., et al.: Current Challenges and Visions in Music Recommender Systems Research. *International Journal of Multimedia Information Retrieval* **7**(2), 95–116 (2018)
32. Strasser, P.: Diversifying search results with maximum marginal relevance. Blog post on Elastic Search Labs (2025)
33. Wu, L., Zheng, Z., Qiu, Z., Wang, H., Gu, H., Shen, T., Qin, C., Zhu, C., Zhu, H., Liu, Q., et al.: A survey on large language models for recommendation. *World Wide Web* **27**(5), 60 (2024)
34. Zhang, Y., Xu, H., Salha-Galvan, G., Han, R., Xiao, F., Huang, Y., Lin, L., Luo, Y., Hu, Y.: Idproxy: Cold-start ctr prediction for ads and recommendation at xiaohongshu with multimodal llms. arXiv preprint arXiv:2603.01590 (2026)
35. Zhao, Z., Fan, W., Li, J., Liu, Y., Mei, X., Wang, Y., Wen, Z., Wang, F., Zhao, X., Tang, J., et al.: Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering* (2024)