Talking to Your Recs: Multimodal Embeddings For Recommendation and Retrieval

Sergio Oramas[1], Andres Ferraro[1], Alvaro Sarasua[1] and Fabien Gouyon[1]

[1]SiriusXM. Oakland, USA

Abstract

Large Language Models (LLMs) excel at understanding complex natural language requests, and even providing recommendations, but they often rely on incomplete or outdated data with respect to platform catalogs. Training or fine-tuning LLMs for these custom catalogs is both costly and challenging. To address this, we propose a method that leverages pre-trained large text embedding models to generate embeddings from catalog descriptions, enriched with multimodal content such as audio or images, as well as collaborative filtering data, using contrastive learning. The resulting enriched embeddings are well-suited for both recommendation and textual search tasks, enabling applications like filtered recommendations, playlist continuation, and playlist generation from text. We evaluate our method through experiments on item recommendation and retrieval, using a real-world music streaming dataset. Our results show substantial improvements in recommendation performance and competitive retrieval performance when compared to off-the-shelf text embeddings and traditional search baselines. We also validate our approach on a public movie dataset, demonstrating its generalizability. Our findings highlight the potential of enhancing language models with additional information and the versatility of our method across diverse domains and applications, all without the need for fine-tuning or training multimodal LLMs from scratch, thereby reducing computational costs.

Keywords

LLMs, Multimodal Recommendation, Retrieval, Content-aware Recommendation

1. Introduction

Large Language Models (LLMs) are quickly expanding to new applications and showing beneficial uses in multiple domains, and specially in music. LLMs showcase an exceptional understanding of language, and are starting to be leveraged in conversational recommender systems and applications such as the so called *AI generated playlists* [1]. However, these large models are trained with general information and therefore need to be adapted when applied in a specific context [2]. The disparity between the knowledge used to train these general-purpose models and the unique internal knowledge and entity catalogues of a company, coupled with the need for continuous knowledge updates, renders these models inadequate for off-the-shelf use in real-world music recommendation applications. Moreover, the expense of fine-tuning or retraining these models to align with an in-house catalog remains significantly high. Approaches like Retrieval Augmented Generation (RAG) [3] present a promising solution to this challenge without the need for model fine-tuning or retraining. This is achieved by conducting a vector search on a corpus of in-house document embeddings. These documents, which usually include textual descriptions of entities, are transformed into embeddings by feeding them into text encoders from large language models optimized for semantic similarity. However, these document embeddings may not fully capture a company's internal knowledge, such as user feedback or content descriptors, making them suitable for retrieval but less effective for recommendations. In this work, we aim to address this issue by combining internal knowledge from various modalities with text embeddings, thereby enhancing the recommendation capabilities of these text embeddings, without the need to retrain or fine-tune the large language models used to generate them.

Adapting LLMs specifically to recommendation tasks has gained recent attention. A review can

MuRS 2024: 2nd Music Recommender Systems Workshop, October 14th, 2024, Bari, Italy

*Corresponding author.

sergio.oramas@siriusxm.com (S. Oramas); andres.ferraro@siriusxm.com (A. Ferraro); alvaro.sarasua@siriusxm.com (A. Sarasua); fabien.gouyon@siriusxm.com (F. Gouyon)

be found in [4]. Examples of recent work include e.g. methods to leverage LLMs for generating descriptions used for recommendation [5], or e.g. an analysis of the performance of prompting LLMs for recommendation compared with state-of-the-art recommenders [6]. However, recent literature suggests that in general recommendation performance of off-the-shelf LLMs is suboptimal, and that further research is needed for their adaptation to state-of-the-art recommendation methods and data [6, 7, 8].

Successful recommendation methods typically leverage multiple modalities of data, either user collaborative data, content features —or both—, areas where LLMs trained on general information may still be deficient. Recommendation with multiple modalities (in particular including content-based methods) has been typically applied with the goal of alleviating cold-start and sparsity issues (e.g. [9, 10]), and also to provide explainable recommendations (e.g. [11]). In particular, the specific domain of music retrieval and recommendation has shown to be a particularly rich playground for exploring the worth of diverse modalities in addition to collaborative filtering (e.g. audio descriptors [12], document similarity [13], or graphs of musical connections [14, 15]). Recent work demonstrated that pre-trained models covering several modalities can be successfully combined for music retrieval and recommendation [16], and that combining different sources and types of data is particularly promising for mitigating current music recommendation limitations [17]. Research in the music domain showed that combining diverse modalities for recommendation can be done in a variety of ways, from simple concatenation [18], to predicting one modality from another [19], or multimodal contrastive learning [20]. Metric learning proved also in other domains to be an effective method to combine heterogenous and complementary information (users' feedback data, audio, image, or text) to improve the quality of the recommendations (e.g. [21, 22, 23]).

In this work, we examine what we believe is a timely and relatively novel hypothesis, that combining text embeddings derived from item descriptions, with data from multiple modalities can improve the performance of those embeddings in recommendation while retaining a comparative performance at retrieval.

In the remainder of this paper, we evaluate an approach that combines text embeddings derived from item descriptions with audio or image, and collaborative filtering data. Item descriptions are created by either inserting tags into a predefined template or by instructing an LLM to generate a description based on these tags —similarly to previous work [24, 25, 26]. Text embeddings are then generated by passing these descriptions through an off-the-shelf pre-trained text embedding model. Embeddings from the different modalities are then integrated into a shared latent space using contrastive learning, and multimodal embeddings are obtained by averaging the projections of the different modality embeddings in this shared space, following the methodology described in [20]. Our evaluation includes a series of experiments involving item recommendation and textual search tasks, employing two datasets: a proprietary dataset from the music domain and an open dataset from the movie domain.

In summary, this work introduces a novel method that enhances pre-trained text embeddings with multimodal content, such as audio, images, and collaborative filtering data, through contrastive learning. This approach significantly improves recommendation performance while maintaining competitive retrieval capabilities, all without the need for fine-tuning or retraining large language models. The method demonstrates robust generalizability across domains, as evidenced by its successful application to both music and movie datasets. Additionally, it opens up versatile possibilities for multimodal retrieval and personalized search, offering a cost-effective solution for improving recommendation systems in diverse contexts.

2. Method

2.1. Multimodal Contrastive Method

The proposed method combines information from diverse item modalities, following a similar methodology to the one described in [20], but changing the tag embeddings encoder for a text embeddings encoder. Our model receives three pre-trained input embeddings, each of them representing a distinct modality, and returns a new multimodal embedding. One of the input embeddings represents item

textual descriptions encoded with a pre-trained text embedding model, another represents user Collaborative Filtering (CF) information obtained through matrix factorization, and the third one represents content features related to the application domain —audio or images— (an illustration is provided in Figure 1, see below for more details on input embeddings computation).

For the proposed model we apply a contrastive learning loss based on InfoNCE [27]. Specifically, we define the contrastive loss between two modalities, $\boldsymbol{\psi}_a$ and $\boldsymbol{\psi}_b$, as:

$\mathcal{L}_{\boldsymbol{\psi}_a,\boldsymbol{\psi}_b} = \sum_{i=1}^{M} -\log \frac{\Xi(\boldsymbol{\psi}_a^i,\boldsymbol{\psi}_b^i,\tau)}{\sum_{k=1}^{2M} \mathbb{1}_{[k\neq i]}\Xi(\boldsymbol{\psi}_a^i,\boldsymbol{\zeta}^k,\tau)}$, where M is the batch size and τ is the temperature parameter.

We define $\Xi(\mathbf{a},\mathbf{b},\tau) = \exp(\cos(\mathbf{a},\mathbf{b})\tau^{-1})$, based on the cosine similarity. $\boldsymbol{\zeta}^k$ is defined as $\boldsymbol{\psi}_a^k$, if $k \leq M$ and else $\boldsymbol{\psi}_b^{k-M}$. This loss function attempts to minimize the distance between the representations of the modalities of the same item while maximizing the distance between any representation of modalities from other items.

We employ three encoders, each dedicated to a specific modality, to generate three representations within our shared space for every item (see 1). Each of the encoders are a simple feed-forward network with one or two dense layers and ReLU activation. During training, we learn the parameters of these encoders by minimizing the cumulative pairwise losses between modalities [1], as in [28]. The objective function, denoted as $\mathcal{L}_{\text{tot}}$, comprises the sum of losses from all pairwise combinations: $\mathcal{L}_{\text{Audio-Text}}$, $\mathcal{L}_{\text{Audio-CF}}$, and $\mathcal{L}_{\text{Text-CF}}$.

2.2. Multimodal embeddings for Recommendation and Retrieval

Once the model is trained with the contrastive method and we want to use it for inference, we obtain the multimodal embedding by averaging the output of each internal encoder for a given item. Following this procedure, we compute a multimodal embedding for every item, and then we can use these embeddings either for recommendation or retrieval.

For recommendation, the multimodal embeddings can be used as item features in any content-based or hybrid recommendation approach. For retrieval, given a text query, we first compute the query embedding by passing the text through a pre-trained text embedding model, followed by projecting it using our model's text encoder. We then use the obtained embedding to perform a nearest neighbour search on the space of all the multimodal item embeddings (see Figure 2).

[1] All the encoders in the evaluated models had 1 hidden layer of 256 units, an output layer of 200 units, dropout of 0.3 and a batch size of 512

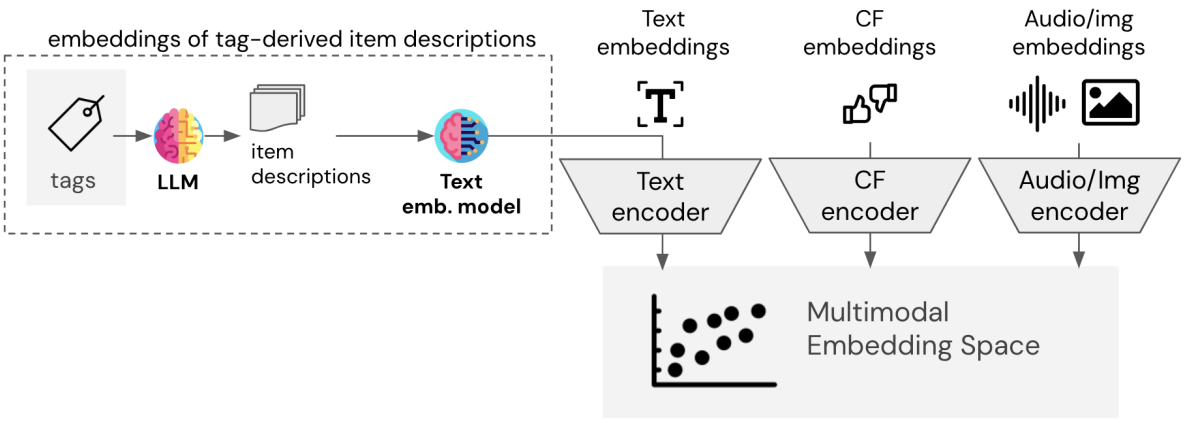

Figure 1: Representation of the internal components used to train the multimodal embeddings.

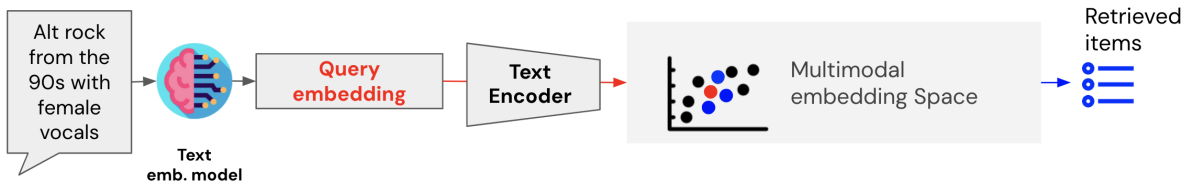

Figure 2: Representation of how our method is used for retrieval with a natural-language query as input.

3. Experimental setup

Our primary focus is to evaluate our method in the music domain, but to test its efficacy across various scenarios and domains, we employ two distinct datasets: a dataset with information about artists collected from a music streaming platform, which contains high-quality single-modality embeddings pre-trained on industry-scale data, and a publicly available dataset from the movies domain. Subsequently, we conduct recommendation and retrieval experiments on each dataset.

3.1. Music dataset and experiments

To evaluate our approach we collected tags and pre-computed collaborative and audio embeddings of artists from a music streaming platform. Collaborative embeddings come from an internal process of matrix factorization of explicit feedback, whereas audio embeddings come from an internal machine learning method [29]. The details about these two collections of embeddings are out of the scope of this paper, we simply use them as input to our method, but they were computed with industry-scale high quality data. The tags of an artist were manually annotated by curators and provides a rich description of its genre, mood, location, decade and musicological characteristics.

We generate two types of descriptions for each music artist using their respective sets of tags. The first type (i.e. $TEMPLATE$) is created by employing a template that organizes the tags into categorized lists of words. A second version of the descriptions (i.e. GEN) is created by tasking an LLM (Claude V3 Haiku) to enhance the semantic expressiveness of the $TEMPLATE$ description. The LLM can also incorporate its own knowledge about the described item, if available (see examples in Table 1). We generated these alternative descriptions to examine whether a text exhibiting a smoother natural language flow can aid the model in any of the tasks.

In the remainder of this work, we refer to the model trained with the contrastive method with a template based on tags as $MULTI_{TEMPLATE}$ and the model trained with descriptions generated from the tags with an LLM as $MULTI_{GEN}$. We also combine template-based descriptions with LLM-based generated descriptions for training the model, and we refer to this as $MULTI_{TEMPLATE+GEN}$. In this latter approach, the model is trained with two embeddings per item: one derived from the $TEMPLATE$ descriptions and the other from the GEN descriptions.

We train our contrastive model with the three modality embeddings —CF, text and audio— of 31,605 artists for the different types of descriptions. As described in Section 2.1, the training objective is a self-supervised objective, not directly related to the two downstream tasks we are evaluating, thus our model is not optimized for any of the tasks.

The recommendation task is evaluated in an item-to-item recommendation scenario. We used for this evaluation the publicly available dataset OLGA [20], used for artist similarity and artist-to-artist recommendation. From this dataset we collected a ground-truth of 6537 artists and a list of positive recommendations for each of the artists. We then evaluate our embeddings by retrieving the k-NN of each artist and comparing them with the ground-truth.

For the retrieval task, we randomly select sets of 1, 2, or 3 tags for which, given a query, there are at least 30 positive results in the dataset. We combine each set of tags with a query template and then utilize an LLM (Anthropic's Claude v3 Haiku model [30]) to enhance the syntactic variability of the

Table 1
Descriptions for the artist "Metallica"

$TEMPLATE$	GEN
The name of this artist is Metallica and is described by the following tags: Genre: speed thrash, pop-metal, metal, classic hard rock. Location: el cerrito, san francisco, california, united states. Decade: nineteen eighty, nineties. Mood: high arousal, negative, aggressive. Lyrics: angry lyrics, clean lyrics, abstract lyrics, depressing lyrics, social lyrics. Instruments: riffing electric guitar, agressive drums, distorted electric guitar, drumset, bass riff, electric guitar, electric guitar solo, rhythm guitar. Musicological: lead vocals, sad harmony, strong backbeat, 4/4 meter. Sonority: electric, high energy music, acoustic, epic, beyond the mainstream, instrumental, extreme aural intensity, outstanding technique, live performance, studio recording. Vocals: gravelly timbre, man, singer, lead vocal.	Metallica is a legendary heavy metal band hailing from the San Francisco Bay Area in California. Blending elements of speed thrash, pop-metal, and classic hard rock, their music is characterized by a high-energy, aggressive sound that has captivated audiences since the 1980s. With a penchant for angry, socially-conscious lyrics delivered in a gravelly, powerful vocal style, Metallica's compositions feature a barrage of riffing electric guitars, thunderous drums, and driving bass lines. The band's musicality is further enhanced by their skilled use of harmony, strong backbeats, and virtuosic instrumental solos. Metallica's sonic palette spans from the electric intensity of their studio recordings to the epic, beyond-the-mainstream experience of their live performances. Whether exploring themes of depression, social unrest, or simply unleashing their raw, unbridled energy, this iconic American metal act has left an indelible mark on the genre and the music world at large.

queries (see Table 2 for some examples). We experimented with various prompts for the LLM until we achieved a set of queries that were both satisfactory and semantically rich. Following this method, we create a dataset of 3000 queries to evaluate the retrieval task. The ground truth for each query comprises items associated with the corresponding tags used to generate that query.

Table 2
Query examples from the music and movies retrieval evaluation datasets

Find me some instrumental guitar music from the 2010s. Can you find me music with an acoustic piano, joyful lyrics, and a catchy chorus?
Can you list films that depict drug abuse and its consequences? I'm in the mood for some Italian films from the 1960s. What do you suggest?

3.2. Movie dataset and experiments

To corroborate our findings in a different domain, we use the user-movie ratings from the Movielens-25M dataset[2] [31] combined with movie tags from the Movielens Tag Genome Dataset 2021[3] [32] and image embeddings released in [23]. These image embeddings were generated from the posters of the movies and produced with the CLIP pre-trained model [33].

The CF embeddings for this dataset are obtained using weighted matrix factorization [34] based on the ratings on Movielens-25M with random-based hyperparameter tuning. Ratings with a value higher than 3 are considered as positive during the evaluation. We reserve 10% of users for evaluation of the recommendation task and use the rest to compute the item CF embeddings.

Text embeddings are created in the same way as in the music dataset, by creating two types of descriptions ($TEMPLATE$ and GEN) using the tags and then feeding them in a pre-trained text embedding model (see examples in Table 3). Then, we train our contrastive model with the three modality embeddings —CF, text and image— of 59,040 items, training one model for each of the item description types.

The recommendation task is evaluated using also a naive k-NN approach. Our intention in both datasets is to compare different features in a recommendation setting and not to compare different recommendation approaches. To generate recommendations, we randomly select 50% of each test user's

[2] https://grouplens.org/datasets/movielens/25m/
[3] https://grouplens.org/datasets/movielens/tag-genome-2021

Table 3
Descriptions for the movie "Toy Story" (shortened for space).

$TEMPLATE$	GEN
Movie: Toy Story. Year: 1995. Directed by: John Lasseter. Starring: Tim Allen, Tom Hanks, Don Rickles, Jim Varney, John Ratzenberger, Wallace Shawn, Laurie Metcalf, John Morris, R. Lee Ermey, Annie Potts. Genres: Adventure, Animation, Children, Comedy, Fantasy. Tags: 3d, action, adventure, alien, animated, animation, [...] time travel, torture, toys, unlikely friendships, very good, violence, witty.	Toy Story is a beloved animated classic that follows the adventures of a group of toys, led by the loyal cowboy doll Woody and the charismatic space ranger Buzz Lightyear. Directed by the visionary John Lasseter, this groundbreaking film from Pixar blends comedy, fantasy, and heartwarming themes as the toys navigate the challenges [...] With a stellar voice cast including Tim Allen [...] a true classic of the animation genre.

positively rated items and use them to compute recommendations. For this, we calculate the average embedding of the selected items and search for the top 200 nearest neighbors. We then compare these recommendations with the remaining items that the users have positively rated.

The retrieval task is evaluated in a manner similar to the music experiment. Using the available tagging data, we created an evaluation dataset consisting of 704 queries following the same methodology.

3.3. Baselines

We compare the performance of our multimodal embeddings against off-the-shelf text embedding models, which enable both recommendation and retrieval tasks. We select the best performing text embedding models available in Hugging Face at the time of the writing of this paper. We tested the models WhereIsAI/UAE-Large-V1 [35] and intfloat/e5-large-v2 [36]. We only report results of the $MULTI$ model trained with WhereIsAI/UAE-Large-V1 embeddings, as it provided the best performance among the two models. In addition, we compare them with the different input modality embeddings —CF, Audio, and Image— only in the recommendation task, as the latter cannot be used directly for retrieval. Finally, we compare the retrieval performance with the standard lexical search baseline BM25 [37].

4. Results and Discussion

Table 4
Recommendation and retrieval performance in terms of nDCG@200. "n/a" depicts methods not directly applicable to the task.

	Music		Movielens	
Model	Recommendation	Retrieval	Recommendation	Retrieval
$MULTI_{TEMPLATE}$	0.5196	0.0759	0.2362	0.0881
$MULTI_{GEN}$	0.5198	0.0815	0.2320	0.0611
$MULTI_{TEMPLATE+GEN}$	**0.5251**	0.1102	**0.2461**	0.0925
$TEXT_{intfloat/e5-large-v2}$	0.3053	**0.1348**	0.1833	0.1116
$TEXT_{WhereIsAI/UAE-Large-V1}$	0.3271	0.1310	0.1226	**0.1551**
CF	0.5082	n/a	0.1545	n/a
$IMG/AUDIO$	0.4439	n/a	0.1563	n/a
$BM25$	n/a	0.0890	n/a	0.0632

4.1. Recommendation

Examining the recommendation performance across both datasets (see Table 4), we observe that the proposed model outperforms the text embedding models and the rest of individual modality embeddings.

We also observe that there are no significant differences between recommendation performance for GEN and $TEMPLATE$ descriptions. However, we notice a slight improvement in recommendation performance by combining GEN and $TEMPLATE$ descriptions of the items when training the $MULTI$ model. Combining multiple descriptions for the items generated with multiple methods seems to work as an up-sampling technique that may suggest potential benefits for learning a better space, as showcased in both datasets. A deeper study is necessary to understand the reasons behind this, but the LLM may have introduced some internal knowledge in the GEN descriptions, which are not present in the $TEMPLATE$ ones, and at the same time, obviate some tags present in the $TEMPLATE$. Therefore, the combination may provide a more complete description of the item. It is noteworthy that achieving a similar or higher performance with the proposed method compared to the CF baseline is particularly relevant. This indicates that the model effectively captures the collaborative information and successfully integrates it with data from other modalities.

4.2. Retrieval

Looking at the results for the retrieval task in both datasets, we observe that the text embedding models perform slightly better than the $MULTI$ models. However, we observe a good performance of the proposed method overall, considering that the proposed model enables natural language queries while improving the recommendation capabilities. The two text embedding models tested show different performance depending on the dataset: intfloat/e5-large-v2 is better in the Music dataset and WereIsAI/UAE-Large-V1 is better in the Movielens dataset. All the embedding-based approaches, including $MULTI$ and the off-the-shelf text embedding models are better than the lexical search baseline BM25. This implies that semantic search in embedding spaces is behaving better than traditional lexical search for complex tag-based queries like those in our evaluation dataset.

We also observe that the proposed method performs closer to the text embedding models (while still slightly worse) in the music domain than in the movies domain. This can be attributed to differences in input feature quality. For instance, the Music dataset includes high-quality manual annotations from music experts, providing more comprehensive descriptions than those created from Movielens tags. Additionally, CF and audio embeddings significantly outperform text embeddings in the music recommendation task, whereas CF and image embeddings are on par with text embeddings in the movie recommendation task. This highlights the disparity in input embedding quality between the two datasets, which seems to be a key factor in the multimodal model's performance.

4.3. Further applications

In this work, we analyzed the characteristics of the embeddings generated by our approach. Results demonstrate its capabilities for text retrieval, while at the same time, showcasing an improved organization of items based on similarity — as evidenced in the k-NN recommendation evaluations. This implies that the top-k results of a text query using our model exhibit higher similarity among items than those retrieved by off-the-shelf text embedding models. Based on these results we see that using our model would be particularly useful for playlist generation from text, where the objective is not only to satisfy a query, but also to provide a coherent listening experience.

In addition, the proposed model offers versatile possibilities beyond text retrieval, supporting queries from various modalities. For instance, it can handle queries using an audio piece, or a combination of audio and text, or a collaborative embedding representing a user profile alongside a text query, thus enabling personalized search.

5. Conclusion

In this study, we introduced a method based on contrastive learning that enhances off-the-shelf text embeddings with multimodal information, avoiding the need for retraining or fine-tuning an LLM. We assessed its effectiveness in recommendation and retrieval tasks across various domains.

In our evaluations, we found that the proposed multimodal embeddings outperform baselines in the *recommendation* task in both music and movies domains, while enabling natural language search —which is not directly possible for collaborative filtering, audio or image features alone. While our model demonstrates competitive performance in *retrieval* tasks, it slightly trails behind off-the-shelf text embeddings. The robust performance of these text embedding models underscores their versatility in various applications and domains, yielding robust item representations beneficial not only for retrieval but also recommendation. However, our findings indicate that their recommendation capabilities significantly improve when integrated with collaborative filtering and content-based embeddings.

Our experiments also demonstrate that combining different types of textual item descriptions enhances performance in both tasks. Exploring the reasons behind this improvement and investigating how additional types of descriptions can further boost performance presents a promising area for future research. Future work could also involve expanding experiments to diverse datasets and domains, comparing with more text embedding models, and exploring diverse modalities. Moreover, this model opens up a myriad of possibilities for playlist creation, multimodal retrieval and personalized search that worth to be explored. Lastly, given the reliance on large internet-trained models, careful analysis of potential biases and risks in recommendations is essential.

References

[1] E. H. Schwartz, Deezer's new ai playlist producer challenges spotify, amazon, youtube music to a dj battle, techradar (2024). URL: https://www.techradar.com/computing/artificial-intelligence/deezers-new-ai-playlist-producer-challenges-spotify-amazon-youtube-music-to-a-dj-battle.

[2] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Barnes, A. Mian, A comprehensive overview of large language models, arXiv preprint arXiv:2307.06435 (2023).

[3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. arXiv:2005.11401.

[4] W. Hua, L. Li, S. Xu, L. Chen, Y. Zhang, Tutorial on large language models for recommendation, in: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 1281–1283. URL: https://doi.org/10.1145/3604915.3609494. doi:10.1145/3604915.3609494.

[5] A. Acharya, B. Singh, N. Onoe, Llm based generation of item-description for recommendation system, in: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 1204–1207. URL: https://doi.org/10.1145/3604915.3610647. doi:10.1145/3604915.3610647.

[6] D. Di Palma, Retrieval-augmented recommender system: Enhancing recommender systems with large language models, in: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 1369–1373. URL: https://doi.org/10.1145/3604915.3608889. doi:10.1145/3604915.3608889.

[7] K. Bao, J. Zhang, Y. Zhang, W. Wang, F. Feng, X. He, Tallrec: An effective and efficient tuning framework to align large language model with recommendation, in: Proceedings of the 17th ACM Conference on Recommender Systems, 2023, pp. 1007–1014.

[8] X. Ren, W. Wei, L. Xia, L. Su, S. Cheng, J. Wang, D. Yin, C. Huang, Representation learning with large language models for recommendation, in: Proceedings of the ACM on Web Conference 2024, 2024. URL: http://dx.doi.org/10.1145/3589334.3645458.

[9] X. Pan, Y. Chen, C. Tian, Z. Lin, J. Wang, H. Hu, W. X. Zhao, Multimodal meta-learning for cold-start sequential recommendation, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 3421–3430. URL: https://doi.org/10.1145/3511808.3557101. doi:10.1145/3511808.3557101.

[10] Y. Zhang, Q. Ai, X. Chen, W. B. Croft, Joint representation learning for top-n recommendation

with heterogeneous information sources, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 1449–1458. URL: https://doi.org/10.1145/3132847.3132892. doi:10.1145/3132847.3132892.

[11] Z. Cai, Z. Cai, Pevae: A hierarchical vae for personalized explainable recommendation., in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 692–702. URL: https://doi.org/10.1145/3477495.3532039. doi:10.1145/3477495.3532039.

[12] T. Pohle, D. Schnitzer, M. Schedl, P. Knees, G. Widmer, On rhythm and general music similarity., in: ISMIR, 2009, pp. 525–530.

[13] M. Schedl, D. Hauger, J. Urbano, Harvesting microblogs for contextual music similarity estimation: a co-occurrence-based framework, Multimedia Systems 20 (2014) 693–705.

[14] S. Oramas, M. Sordo, L. Espinosa-Anke, X. Serra, A semantic-based approach for artist similarity, in: ISMIR, 2015.

[15] F. Korzeniowski, S. Oramas, F. Gouyon, Artist similarity for everyone: A graph neural network approach, Transactions of the International Society for Music Information Retrieval 5 (2022).

[16] M. Won, S. Oramas, O. Nieto, F. Gouyon, X. Serra, Multimodal metric learning for tag-based music retrieval, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 591–595.

[17] A. Ferraro, Music cold-start and long-tail recommendation: Bias in deep representations, in: Proceedings of the 13th ACM Conference on Recommender Systems, 2019, p. 586–590. URL: https://doi.org/10.1145/3298689.3347052.

[18] S. Oramas, O. Nieto, F. Barbieri, X. Serra, Multi-label music genre classification from audio, text, and images using deep features, in: ISMIR, 2017.

[19] A. Van den Oord, S. Dieleman, B. Schrauwen, Deep content-based music recommendation, Advances in neural information processing systems 26 (2013).

[20] A. Ferraro, J. Kim, S. Oramas, A. Ehmann, F. Gouyon, Contrastive learning for cross-modal artist retrieval, in: ISMIR, 2023. URL: https://arxiv.org/abs/2308.06556.

[21] F. Liu, Z. Cheng, C. Sun, Y. Wang, L. Nie, M. Kankanhalli, User diverse preference modeling by multimodal attentive metric learning, in: Proceedings of the 27th ACM International Conference on Multimedia, MM '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 1526–1534. URL: https://doi.org/10.1145/3343031.3350953. doi:10.1145/3343031.3350953.

[22] C.-K. Hsieh, L. Yang, Y. Cui, T.-Y. Lin, S. Belongie, D. Estrin, Collaborative metric learning, in: Proceedings of the 26th International Conference on World Wide Web, WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2017, p. 193–201. URL: https://doi.org/10.1145/3038912.3052639. doi:10.1145/3038912.3052639.

[23] I. Avas, L. Allein, K. Laenen, M.-F. Moens, Align macridvae: Multimodal alignment for disentangled recommendations, in: European Conference on Information Retrieval, Springer, 2024, pp. 73–89.

[24] S. Doh, K. Choi, J. Lee, J. Nam, Lp-musiccaps: Llm-based pseudo music captioning, in: ISMIR, 2023. URL: https://arxiv.org/abs/2307.16372.

[25] D. McKee, J. Salamon, J. Sivic, B. Russell, Language-guided music recommendation for video via prompt analogies, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14784–14793.

[26] J. P. Gardner, S. Durand, D. Stoller, R. M. Bittner, Llark: A multimodal instruction-following language model for music, in: Forty-first International Conference on Machine Learning, 2023.

[27] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).

[28] A. Ferraro, X. Favory, K. Drossos, Y. Kim, D. Bogdanov, Enriched music representations with multiple cross-modal contrastive learning, IEEE Signal Processing Letters 28 (2021) 733–737.

[29] M. C. McCallum, F. Korzeniowski, S. Oramas, F. Gouyon, A. F. Ehmann, Supervised and unsupervised learning of audio representations for music understanding, in: ISMIR, 2022.

[30] Anthropic, https://paperswithcode.com/paper/the-claude-3-model-family-opus-

sonnet-haiku, Papers With Code (2024). URL: https://paperswithcode.com/paper/the-claude-3-model-family-opus-sonnet-haiku.

[31] F. M. Harper, J. A. Konstan, The movielens datasets: History and context, ACM Trans. Interact. Intell. Syst. 5 (2015). URL: https://doi.org/10.1145/2827872. doi:10.1145/2827872.

[32] J. Vig, S. Sen, J. Riedl, The tag genome: Encoding community knowledge to support novel interaction, ACM Trans. Interact. Intell. Syst. 2 (2012). URL: https://doi.org/10.1145/2362394.2362395. doi:10.1145/2362394.2362395.

[33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[34] Y. Hu, Y. Koren, C. Volinsky, Collaborative filtering for implicit feedback datasets, in: Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), 2008, pp. 263–272.

[35] X. Li, J. Li, Angle-optimized text embeddings, arXiv preprint arXiv:2309.12871 (2023).

[36] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, F. Wei, Text embeddings by weakly-supervised contrastive pre-training, arXiv preprint arXiv:2212.03533 (2022).

[37] S. Robertson, H. Zaragoza, et al., The probabilistic relevance framework: Bm25 and beyond, Foundations and Trends® in Information Retrieval 3 (2009) 333–389.