Harnessing High-Level Song Descriptors towards Natural Language-Based Music Recommendation

Elena V. Epure, Gabriel Meseguer Brocal, Darius Afchar, Romain Hennequin
Deezer Research, Paris, France
research@deezer.com

Abstract

Recommender systems relying on Language Models (LMs) have gained popularity in assisting users to navigate large catalogs. LMs often exploit item high-level descriptors, i.e. categories or consumption contexts, from training data or user preferences. This has been proven effective in domains like movies or products. However, in the music domain, understanding how effectively LMs utilize song descriptors for natural language-based music recommendation is relatively limited. In this paper, we assess LMs effectiveness in recommending songs based on user natural language descriptions and items with descriptors like genres, moods, and listening contexts. We formulate the recommendation task as a dense retrieval problem and assess LMs as they become increasingly familiar with data pertinent to the task and domain. Our findings reveal improved performance as LMs are fine-tuned for general language similarity, information retrieval, and mapping longer descriptions to shorter, high-level descriptors in music.

1 Introduction

Music recommender systems are often used to assist users in navigating the vast catalogs offered by streaming platforms. Recently, Language Models (LMs) have demonstrated efficacy in recommending items such as songs, movies, or books (Sanner et al., 2023; Penha and Hauff, 2020). They achieve this by 1) matching concise user profiles derived from item preferences expressed in natural language against the available set of items; and 2) drawing upon their knowledge about specific items gained during pre-training or after fine-tuning for the specific task or domain.

A common thread among these strategies is the reliance of LMs on item high-level descriptors, such as genres (e.g. "romantic comedy" for a movie) or consumption contexts (e.g. "running" for music), within the pre-training corpus or during

user preference elicitation phase. Empirical evidence suggests that these descriptors play a crucial role in generating relevant recommendations with LMs, in cases like near cold-start or exploration in movie or product domains (Sanner et al., 2023; López et al., 2021; Malkiel et al., 2020). Leveraging these descriptors has been proven particularly valuable and could surpass the importance of item examples alone such as in collaborative filtering (Penha and Hauff, 2020; Sanner et al., 2023).

In music however, our understanding of how efficiently LMs utilize item textual features and user preference description for music recommendation is comparatively limited. Previous research has emphasized the significance of natural language features such as tags in improving music retrieval (Doh et al., 2023b; Wu* et al., 2023) and captioning algorithms (Gabbolini et al., 2022; Doh et al., 2023a). These accessible descriptors help bridge the semantic gap between audio and more complex song descriptions provided by humans (Celma Herrada et al., 2006). Yet, such a study is lacking in the field of music recommendation despite advancements in related areas like conversational recommendation systems (Chaganty et al., 2023; Jannach et al., 2021; Gupta et al., 2023; Jin et al., 2019).

In this work, we study the efficacy of LMs for song recommendation when users express their preferences in natural language and items are associated with high-level descriptors such as music genres, styles, moods, and listening contexts. As no large dataset linking natural language user preferences to music descriptors is available, we re-purpose an existing dataset originally created for music captioning (Doh et al., 2023a). Subsequently, we formulate the recommendation task as a dense retrieval problem (Wang et al., 2022) and propose evaluating the LMs as they become increasingly familiar with data pertinent to the target task (recommendation as retrieval) and domain (music). This contrasts with previous approaches

to integrating LMs, that typically involve encoding low-level item data (such as audio or embedded metadata) and user queries (Doh et al., 2023b; Wu* et al., 2023), and are likely less effective because of the inherent semantic gap in music.

Our results show that a) pre-trained LMs with no further fine-tuning perform quite poorly on this task in music; b) the performance gets better when the LMs are progressively fine-tuned for text similarity, for multi-domain query retrieval and for mapping longer descriptions on shorter, high-level descriptors in music. Our approach renders the dataset and models suitable also for retrieving explanations based on song high-level descriptors or encoding any type of item textual information in multi-modal music systems. We release the code and fine-tuned models at `https://github.com/deezer/nlp4musa_melscribe`.

2 Related Work

Pre-trained LMs are widespread and their usage spans various tasks, including generating synthetic data, producing system utterances, and providing recommendations. Penha and Hauff (2020) fine-tune BERT on recommendation conversations on movies, books, or music extracted from Reddit and assess the resulting model's capability to retrieve the most relevant utterance containing a suitable recommendation from existing conversations. Similarly, Chaganty et al. (2023) model the recommendation task as retrieval and fine-tune multiple checkpoints of a T5 encoder to create a common embedding space of conversations and song metadata such (i.e. title, artist and album). Mysore et al. (2023) build a new dataset by using Instruct-GPT to generate narrative-driven recommendations in the Point-of-Interest (PoI) domain. Then, they fine-tune dense retrieval models for this domain.

Other studies depend on LMs for generating system inputs in natural language. Hayati et al. (2020) build a dialogue model comprising two distinct language modules tailored to the recommendation seeker and the recommender system. Kostric et al. (2021) fine-tune a T5 for learning to generate relevant indirect questions about the context of item consumption in order to help the user to elicit preferences when these are not clearly defined.

Similar to these works, we adopt a recommendation as retrieval approach and assess LMs for our song recommendation task in scenarios where music preferences are given in natural language. How-

ever, our recommendation module relies solely on high-level descriptors rather than low-level ones extracted from audio or embedded metadata (e.g. song titles or artists). Our hypothesis is that as the model needs to bridge a narrower semantic gap between these two music information sources, it would yield improved outcomes.

3 Method

The recommendation task we consider is the learning to rank setup: given r, a recommendation request in natural language, and $\mathcal{S}$, a corpus of N songs where each song s has associated $T_s, |T_s| \geq 1$, a set of high-level descriptors (words or short phrases describing music), the recommendation task is addressed with a ranking function ρ over the collection $\mathcal{S}$. If we consider an encoder LM with an embedding function f, which takes as input text and outputs a vector, then the ranking could be computed based on the dot-product similarity between the embeddings of the request r and the concatenated, and alphabetically sorted, high-level descriptors of each song T_s: $\rho(r, \mathcal{S}) = (f(r)^\top f(\text{concat}(T_{s_1})), ..., f(r)^\top f(\text{concat}(T_{s_N})))$. The problem revolves around having an efficient text encoder f in music for embedding high-level descriptors and recommendation requests.

For this, we train a music text bi-encoder using the Generative Pseudo-labeling method (GPL) (Wang et al., 2022). *Bi-encoders* (Reimers and Gurevych, 2019) rely on two siamese encoders, comprising a pre-trained LM such as BERT (Devlin et al., 2019) followed by a pooling layer (e.g. mean). To adapt it to music, we fine-tune it using contrastive learning, which entails that similar input texts (song high-level descriptors on one side and natural language recommendation requests on the other side) are embedded close together in the vector space, while dissimilar ones are far apart.

Two steps are essential in this method: *the automatic negative mining* and *the pseudo-labeling of training examples*. We mine hard negatives (r, T_s^-) using pre-trained bi-encoders as follows: for each query r we retrieve the top K most similar high-level descriptors, $T_k, k \neq s$. However, sometimes negatives could be closer to an actual positive example or even a false negative. For example, there might be instances when a particular user request for recommendation matches other songs as well, but this was not explicitly labeled in the training data. As in GPL (Wang et al., 2022),

we resort to soft-labeling the training data, instead of considering a negative as a true negative. For each tuple (r, T_s, T_s^-), we compute a margin score $\delta_s = g(r, \text{concat}(T_s)) - g(r, \text{concat}(T_s^-))$ with a cross-encoder g as a teacher. Compared to bi-encoders that separately encode input texts to a shared vector space, *cross-encoders* take as input the concatenated texts and produce as output a similarity score. Cross-encoders are known to be more effective at the text similarity task than bi-encoders but do not scale well (Thakur et al., 2021).

Implementation Details We choose models pre-trained on `ms-marco`, a dataset with search queries and documents from various domains (Chen et al., 2015), because of its similarity to our domain and task, that proved useful experimentally too. We fine-tune `msmarco-bert-base-dot-v5`[1] on our music data for 1 epoch and $140K$ steps with a batch size of 4. This model is chosen as the backbone of our bi-encoder as it yields the best performance in our experiments. For each r, we use `msmarco-distilbert-base-v3` and `msmarco-MiniLM-L-6-v3` to mine 30 negative examples. To soft-label the training data, we fine-tune a domain-specific cross-encoder on MusicCaps (Agostinelli et al., 2023), a dataset with human-created song captions and descriptors.

Other Baselines BERT (Devlin et al., 2019) and MPNET (Song et al., 2020) are pre-trained on general language corpora using various objectives such as masked language modeling or permuted language. We apply mean as pooling function to all token embeddings to derive a fixed-size embedding for the given input. Then, we consider the best-performing bi-encoders reported on `sentence-transformers` (Reimers and Gurevych, 2019). Other baselines we include are text encoders obtained from multi-modal (audio-text) representation learning (Wu* et al., 2023; Doh et al., 2023b). Their text encoding branch is initialized with BERT or RoBERTa weights. One difference between TTMR (Doh et al., 2023b) and CLAP (Wu* et al., 2023) lies in the training dataset and the audio encoding branch. A `tf-idf` sparse representation is also considered in the experiments. Although such a text encoding does not generalise to new vocabulary, we expect it to work well when recommendation requests and music descriptors have high exact term overlap.

---

[1]The name of the used pre-trained models reflects the training dataset (`msmarco`), the base text encoder (`bert-base`) and the text similarity function used in training (`dot`)

4 Datasets and Evaluation Details

Since there is no large dataset linking natural language user preferences (r) with high-level song descriptors (T_s) in music recommendation, we re-purpose a dataset originally created for music captioning, LP-MusicCaps (Doh et al., 2023a), to fine-tune our music text bi-encoder. LP-MusicCaps was created from three pre-existing datasets of audio tracks annotated with tags (the ECALS subset of the Million Song Dataset (MSD) (Doh et al., 2023b), MusicCaps (MC) (Agostinelli et al., 2023), and Magnatagtune (MTT) (Law et al., 2009)) by using an instruction-based LLM to generate captions from the given high-level descriptors.

Our goal is to extract training pairs (r, T_s) from LP-MusicCaps by ensuring the compatibility with the desired use case of the text encoder, for conversational music recommendation. Compared to narrative-driven recommendations, user requests are unlikely *long* in synchronous conversations (Chaganty et al., 2023; Mysore et al., 2023). As the length of each caption in LP-MusicCaps ranges from a sentence to a paragraph, we first split paragraphs in sentences. Then, as in GPL where each query is seen with multiple documents during training, we ensure multiple sets of high-level descriptors for each sentence. We sample up to 3 variations of high-level descriptors from the original T_s: first from overlapping high-level descriptors (i.e. tags or phrases from T_s *found* in the sentence) and then from the non-overlapping ones (i.e. tags or phrases from T_s *not found* in the sentence). Like this we simulate cases where there is a varying number of high-level descriptors per song, and some may be irrelevant to the description.

In conversational requests for music recommendations, we could find high-level descriptors T_s similar to those in LP-MusicCaps (Chaganty et al., 2023). However, the user utterance r might be formatted like a *request* instead of a statement as in a song caption (Jannach et al., 2021; Chaganty et al., 2023). While we could rephrase each sentence in `train` as a request, this step is costly and might prove unnecessary. The semantic similarity between the two forms of the song description (request versus statement) and the same set of high-level descriptors, T_s, might be comparable as it likely relies on topical cues. In order to check our hypothesis, we rephrase the single-sentence captions from MusicCaps `test` split from statements into requests for music recommendation

	Tf-Idf	CLAP$_{text}$	TTMR$_{text}$	BERT	all-MiniLM	msmarco-BERT	Ours
MTT	57.7 ± 0.8	13.5 ± 0.3	7.8 ± 0.6	4.8 ± 0.4	33.3 ± 0.6	32.1 ± 0.2	$\mathbf{62.8 \pm 0.5}$
MSD	30.6 ± 2.3	3.4 ± 0.1	5.1 ± 0.1	4.5 ± 0.0	19.5 ± 0.2	20.7 ± 0.1	$\mathbf{47.9 \pm 0.3}$
MC	$\mathbf{89.4 \pm 0.4}$	36.5 ± 1.1	19.9 ± 0.2	24.3 ± 0.9	59.9 ± 0.9	66.1 ± 0.2	84.8 ± 0.2
MC$_{reco}$	$\mathbf{77.7 \pm 0.5}$	27.6 ± 0.4	17.9 ± 1.2	16.3 ± 0.1	48.3 ± 0.4	50.7 ± 1.0	70.1 ± 0.4

Table 1: Recall@10 (mean $\pm$ std) of the all baselines on the LP-MusicCaps test splits.

with Llama3[2] (more details in Appendix A).

We use only the MSD and MTT `train` splits of LP-MusicCaps for training and keep MC completely unseen. Each song s is associated with a concatenated set of high-level descriptors T_s. The mapping from the description r to the most likely $T_k, k \in S$ is the proxy for recommending the song k. Then, we compute Recall@10 at the level of descriptors (T_s), and not at the song level (s) as multiple songs could have the same set of high-level descriptors T_s and T_s is the only song information that we consider in this work. For each dataset, we produce 3 test sets by sampling a different set of high-level descriptors per song s and description r. We then report mean and standard deviation (std).

5 Results

Table 1 presents the results obtained from the evaluation of the proposed music text encoder (`Ours`) and the baselines on song recommendation. We could notice that `tf-idf` is a strong encoding function on these datasets where there is a large exact term overlap between the song description r and the high-level descriptors T_s. Though, on the MTT and MSD datasets where this happens less frequently (see Table 2 for the ratio of descriptor words found in the song description), `tf-idf`, although competitive, falls short. Pre-trained LMs such as BERT achieve poor results, most likely because high-level descriptors being short have insufficient context to derive meaningful embeddings[3]. Although, when used as part of bi-encoders and fine-tuned on a relevant text similarity task (information retrieval), we could see the performance increasing: `msmarco-BERT` is the best dense retrieval model from the `sentence-transformers` collection; we also report results for the second-best, `all-MiniLM`. Similarly, fine-tuning BERT (or variants) on text-audio similarity seem to lead to better text embeddings (see scores for CLAP$_{text}$ and TTMR$_{text}$). Yet, their performance is less good when exploiting high-level descriptors instead of audio.

test	#Requests	#Descriptors	Shared Words
MTT	4462	188	0.15
MSD	34631	1054	0.23
MC	2357	6930	0.41
MC$_{reco}$	2357	6930	0.34

Table 2: Number of requests, unique descriptors, and mean ratio of shared words between each pair (r, T_s) in `test`. It could be noticed that in the MC dataset, 40% of descriptor words are found in the description / request.

The fine-tuned bi-encoder achieves significantly higher scores than all the other dense retrievers. Compared to $tf - idf$, it does not depend on a pre-established vocabulary. Thus, by design, it should generalise to new high-level descriptors and is more robust to synonyms and language variations.

Baselines' scores on the rephrased MC test set are lower compared to those on the original MC dataset. Manual checks revealed that rephrasing descriptions as requests sometimes omitted initial descriptors, thus making it difficult to distinguish between the effects of rephrasing and information loss (a couple of examples are shown in Table 4 in Appendix A). Finally, a more detailed qualitative analysis is presented in Appendix B.

6 Conclusion

Conversational recommender systems or interfaces based on natural language have emerged as a practical alternative to dynamically elicit preferences from the users in cold-start or exploration cases. LMs have emerged as central to these systems. In this work, we analysed the efficacy of LMs towards song recommendation when users express their preferences in natural language and items have high-level descriptors. We showed that a bi-encoder fine-tuned in multiple phases first for the task and then for the domain is quite competitive. Future works aims at improving the model on out-of-distribution data, integrating more specialized music knowledge and personalisation during its fine-tuning, and joining the proposed encoder with other modalities for song recommendation.

---

[2] https://ai.meta.com/blog/meta-llama-3/
[3] Similar results were obtained for `MPNet`.

7 Ethical Considerations

The fine-tuned models, which will be released, target only English-language content and have been exposed primarily to music descriptions and descriptors that mostly refer to Western-centered music, with a limited number of music descriptions and descriptor set pairs. Additionally, we are aware that music descriptors such as mood or genre can be specific to individuals, groups, or cultures. However, the embeddings we obtain with the fine-tuned models are deterministic and do not take into account any form of localization or personalization, which is a limitation with ethical implications.

References

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. 2023. Musiclm: Generating music from text. *Preprint*, arXiv:2301.11325.

Òscar Celma Herrada, Herrera Boyer, Xavier Serra, et al. 2006. Bridging the music semantic gap. In *Proceedings of the Workshop on Mastering the Gap, From Information Extraction to Semantic Representation, held in conjunction with the European Semantic Web Conference; 2006 Jun 11-14; Budva, Montenegro.[Aachen]: CEUR Workshop Proceedings; 2006.* CEUR Workshop Proceedings.

Arun Tejasvi Chaganty, Megan Leszczynski, Shu Zhang, Ravi Ganti, Krisztian Balog, and Filip Radlinski. 2023. Beyond single items: Exploring user preferences in item sets with the conversational playlist curation dataset. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2754–2764, New York, NY, USA. Association for Computing Machinery.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *Preprint*, arXiv:1504.00325.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023a. Lp-musiccaps: Llm-based pseudo music captioning. In *Proceedings of the 24th International Society for Music Information Retrieval Conference*.

SeungHeon Doh, Minz Won, Keunwoo Choi, and Juhan Nam. 2023b. Toward universal text-to-music retrieval. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

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*, pages 11401–11415, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Raghav Gupta, Renat Aksitov, Samrat Phatale, Simral Chaudhary, Harrison Lee, and Abhinav Rastogi. 2023. Conversational recommendation as retrieval: A simple, strong baseline. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 155–160, Toronto, Canada. Association for Computational Linguistics.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. INSPIRED: Toward sociable recommendation dialog systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8142–8152, Online. Association for Computational Linguistics.

Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Comput. Surv.*, 54(5).

Yucheng Jin, Wanling Cai, Li Chen, Nyi Nyi Htun, and Katrien Verbert. 2019. Musicbot: Evaluating critiquing-based music recommenders with conversational interaction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 951–960, New York, NY, USA. Association for Computing Machinery.

Ivica Kostric, Krisztian Balog, and Filip Radlinski. 2021. Soliciting user preferences in conversational recommender systems via usage-related questions. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, pages 724–729, New York, NY, USA. Association for Computing Machinery.

Edith Law, Kris West, Michael I Mandel, Mert Bay, and J Stephen Downie. 2009. Evaluation of algorithms using games: The case of music tagging. In *ISMIR*, pages 387–392.

Federico López, Martin Scholz, Jessica Yung, Marie Pellat, Michael Strube, and Lucas Dixon. 2021. Augmenting the user-item graph with textual similarity models. *arXiv preprint arXiv:2109.09358*.

Itzik Malkiel, Oren Barkan, Avi Caciularu, Noam Razin, Ori Katz, and Noam Koenigstein. 2020. RecoBERT: A catalog language model for text-based recommendations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1704–1714, Online. Association for Computational Linguistics.

Sheshera Mysore, Andrew Mccallum, and Hamed Zamani. 2023. Large language model augmented narrative driven recommendations. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, pages 777–783, New York, NY, USA. Association for Computing Machinery.

Gustavo Penha and Claudia Hauff. 2020. What does bert know about books, movies and music? probing bert for conversational recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*, RecSys '20, page 388–397, New York, NY, USA. Association for Computing Machinery.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Scott Sanner, Krisztian Balog, Filip Radlinski, Ben Wedin, and Lucas Dixon. 2023. Large language models are competitive near cold-start recommenders for language- and item-based preferences. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, page 890–896, New York, NY, USA. Association for Computing Machinery.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2345–2360, Seattle, United States. Association for Computational Linguistics.

Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*.

A Rephrased Song Descriptions

In order to rephrase the single-sentence song descriptions from the MC dataset, we rely on `meta-llama/Meta-Llama-3-8B-Instruct`, which we prompt as follows:

> *Rephrase this as a music recommendation request from a user: «original song description»*
> *Do not use greetings, thanks or emojis.*
> *Keep it short, preferably single-sentence.*
> *Output:*

Each model generation request has a randomly initialised temperature from the list of values {0.80, 0.85, 0.90, 0.95}; randomly initialised top_p value from the same list of values as the temperature; and a randomly initialised seed with a value between 0 and 99999999. Examples are given in Table 3.

Examples of information loss regarding the descriptors when rephrasing song descriptions as requests are presented in Table 4.

B Qualitative Analysis

In Table 5, we present multiple examples of song descriptions from the LP-MusicCaps data, together with their ground-truth associated song descriptors and top five predictions by our model. In the first row, none of the top 5 sets of descriptors matches exactly the ground-truth; however, most retrieved descriptor sets are still relevant. In the next examples, the ground-truth descriptor set is found among those retrieved by the model. However, we could also notice that sometimes irrelevant descriptor sets are returned such as "soft piano" in the forth example, or "exotic" and "amateur recording, r&b" in the last example.

LP-MusicCaps Song Description	Equivalent Recommendation Request
This amateur recording features a steeldrum melody in a higher register creating a joyful and tropical atmosphere.	I'd love to hear some more upbeat tropical music with a similar steeldrum melody in a higher register.
This is a beautiful folk song, embodying the traditional feel of a middle eastern song, featuring a powerful male voice accompanied by the rhythmic beats of the darbuka and the haunting melody of the oud instrument at a moderate tempo, surrounded by other enchanting middle eastern instruments.	I'm looking for music that sounds like a traditional Middle Eastern folk song with a powerful male vocalist accompanied by darbuka and oud at a moderate tempo.
This fingerstyle-guitar track features delicate acoustic guitar melodies played at a medium tempo, creating a classical atmosphere that is both emotional and soothing.	Recommend a fingerstyle guitar track with a classical atmosphere, featuring delicate acoustic guitar melodies played at a medium tempo, perfect for evoking emotions and providing a soothing listening experience.
This k-pop love song features a male vocalist singing in Korean with a youthfully sentimental tone, set to a melodic and dulcet medium tempo track infused with world music influences, including atmospheric synths and chimes, a romantic piano, steady drumming and straightforward bass lines, all backed by a boy band chorus for a pleasantly emotional and ambient experience.	Can you recommend a K-pop love song with a youthful sentimental tone, featuring a male vocalist singing in Korean over a melodic medium tempo track with atmospheric synths, chimes, piano, and boy band chorus?

Table 3: LP-MusicCaps song descriptions and the equivalent request for music recommendation rephrased with the `meta-llama/Meta-Llama-3-8B-Instruct` model.

Original Song Description	Rephrased Song Request	Music Descriptors
This amateur recording features a steeldrum melody in a higher register creating a joyful and tropical atmosphere.	I'd love to hear some more upbeat tropical music with a similar steeldrum melody in a higher register.	steeldrum, higher register, amateur recording
This cinematic masterpiece features a blend of haunting sound effects and triumphant horn honking that transports the listener on a thrilling journey through soundscapes.	I'm looking for music that blends haunting sound effects with triumphant horn honking to create an immersive and thrilling soundscape.	cinematic, soundeffects, horn honking

Table 4: Underlined descriptors are no longer mentioned in the rephrased form of song descriptions as requests.

Song Description / Request	Music Descriptors (Ground-truth)	Top 5 Predictions by Ours
This heartfelt ballad showcases a soulful and sad low-quality sustained strings melody intertwined with a mellow piano melody, and a soft female vocal, resulting in an emotionally charged and sonically rich experience for listeners.	low quality, sad, sustained strings melody, ballad, mellow piano melody	[1] soft piano [2] low quality, emotional female vocal, mellow piano melody, live performance, r&b [3] emotional, low quality, reverberant female vocal, sad acoustic rhythm guitar chord progression, soft rock [4] soft [5] sad
This pop song features a captivating teen female vocal delivering melodic singing over an acoustic guitar and simple drum track that evoke a melancholic, emotional vibe.	acoustic guitar, emotional, teen female vocal, melodic singing, simple drum track, pop music	[1] *acoustic guitar, emotional, teen female vocal, melodic singing, simple drum track, pop music* [2] acoustic guitar [3] acoustic drums [4] bass drum [5] pop, acoustic rhythm guitar, quiet playback, resonant, heartfelt, noisy, emotional, passionate female vocal
This song is a perfect blend of country and pop with a touching singer-songwriter flair, featuring an emotional, soulful male voice that's accompanied by the soft strums of an acoustic guitar.	emotional, male singer, country / pop / singersongwriter	[1] soft [2] acoustic guitar [3] *emotional, male singer, country / pop / singersongwriter* [4] male voice [5] soft piano
Experience an otherworldly journey through an amateur recording filled with out-of-this-world digital sounds, a hair-raising riser, and a hauntingly atmospheric vibe.	atmospheric	[1] amateur recording [2] *atmospheric* [3] digital sound effects, amateur recording [4] exotic [5] amateur recording, r&b

Table 5: Song descriptions, their ground-truth descriptor sets, and top 5 predicted descriptor sets by Ours.