NLP4MusA 2024

Proceedings of the 3rd Workshop on NLP for Music and Audio (NLP4MusA)

15 November, 2024
Oakland, USA

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

Introduction

Welcome to the 3rd Workshop on NLP for Music and Audio (NLP4MusA)! NLP4MusA aims to bring together researchers from various disciplines related to music and audio content on one hand, and NLP on the other. It focuses on topics such as:

- NLP approaches applied to music analysis and generation
- Lyrics analysis and generation
- Exploiting music related texts in music recommendation
- Use of LLMs in music and spoken audio contexts
- Taxonomy learning
- Podcasts recommendations
- Music captioning
- Multimodal representations

The workshop is held in Oakland (CA), in conjunction with ISMIR 2024, and spans half a day featuring a keynote followed by presentations of the accepted papers through short talks and a poster session. The accepted papers cover topics of high relevance to the intersection of music, audio and NLP, including the use of large language models for music tasks such as recommendation, information extraction and lyrics analysis. They also explore the creation and use of new datasets for training and evaluating these models, as well as the generalization capabilities of these models across languages, musical genres and cultural contexts.

We are honored to have Noah Smith (University of Washington, Allen Institute for Artificial Intelligence) as our keynote speaker for this edition of the workshop. His talk explores the challenge of building a language model for MIR. We include the abstract of his talks in this volume.

In response to our call for papers, we received 33 submissions. Each submission was rigorously reviewed by two Program Committee members selected for their expertise. Based on the reviewers' feedback, we accepted 18 papers (55%).

We are extremely grateful to the authors for their valuable contributions and to the Programme Committee members for their detailed and helpful reviews. We also thank our sponsor, Deezer, and our host SiriusXM, who have helped making the workshop possible in this form. We hope you find the workshop insightful and inspiring!

Anna Kruspe, Sergio Oramas, Elena V. Epure, Mohamed Sordo, Benno Weck, SeungHeon Doh, Minz Won, Ilaria Manco, Gabriel Meseguer Brocal

November 2024

Organisers:

Anna Kruspe (Munich University of Applied Sciences)

Sergio Oramas (SiriusXM)

Elena V. Epure (Deezer)

Mohamed Sordo (SiriusXM)

Benno Weck (UPF)

SeungHeon Doh (KAIST)

Minz Won (Suno)

Ilaria Manco (QMUL)

Gabriel Meseguer Brocal (Deezer)

Program Committee:

Morteza Behrooz (Worcester Polytechnic Institute)

Dimitrios Bralios (University of Illinois Urbana-Champaign)

Jonah Casebeer (Adobe Research)

Keunwoo Choi (Genentech)

Shuqi Dai (Carnegie Mellon University)

Matthew Davies (SiriusXM)

Hao-Wen Dong (University of Michigan)

Andres Ferraro (SiriusXM)

Masataka Goto (AIST)

Haven Kim (University of California San Diego)

Junghyun Koo (Seoul National University)

Jongpil Lee (Neutune)

Junwon Lee (KAIST)

Kongmeng Liew (University of Canterbury)

Pasquale Lisena (EURECOM)

Manuel Moussallam (Deezer)

Zachary Novack (University of California San Diego)

Julian Parker (Native Instruments)

Lorenzo Porcaro (Joint Research Centre, European Commission)

Bruno Sguerra (Deezer)

Rosa Stern (Sonos)

Pragati Verma (SiriusXM)

Kento Watanabe (AIST)

Shuo Zhang (Georgetown University)

Yixiao Zhang (QMUL)

Ge Zhu (University of Rochester)

Invited Speaker:

Noah Smith (University of Washington, Allen Institute for Artificial Intelligence)

Invited Talk

Noah Smith: Imagining a Music Language Model

Language models and their multimodal variants now present many exciting new opportunities for advancing music processing applications. As a researcher in natural language processing for over twenty-five years, and as a musician for even longer, I've recently started learning about music IR and related challenges with some of my students. In this talk, I'll offer some opinionated observations, technical ideas, and lessons learned from NLP that I believe could be useful for the music processing community. These include matters of evaluation methodology, the roles of data and theory, the framing of problems, and guiding questions about who we are building technology for. I'll also reflect a bit on our efforts to improve the state of the art in open (multimodal) language models and speculate about a grand challenge: building a (hopefully open) language model for MIR.

Table of Contents

Genre-Conformity in the Topics of Lyrics and Song Popularity 1
Anna Aljanaki

PIAST: A Multimodal Piano Dataset with Audio, Symbolic and Text 5
Hayeon Bang, Eunjin Choi, Megan Finch, Seungheon Doh, Seolhee Lee, Gyeong-Hoon Lee and Juhan Nam

Lyrics Transcription in Western Classical Music with Whisper: A Case Study on Schubert's Winterreise 11
Hans-Ulrich Berendes, Simon Schwär and Meinard Müller

Harnessing High-Level Song Descriptors towards Natural Language-Based Music Recommendation 17
Elena V. Epure, Gabriel Meseguer Brocal, Darius Afchar and Romain Hennequin

NLP Analysis of Environmental Themes in Phish Lyrics Across Concert Locations and Years . . . 25
Anna Farzindar and Jason Jarvis

A Retrieval Augmented Approach for Text-to-Music Generation 31
Robie Gonzales and Frank Rudzicz

Information Extraction of Music Entities in Conversational Music Queries 37
Simon Hachmeier and Robert Jäschke

Leveraging User-Generated Metadata of Online Videos for Cover Song Identification 43
Simon Hachmeier and Robert Jäschke

Can Impressions of Music be Extracted from Thumbnail Images? 49
Takashi Harada, Takehiro Motomitsu, Katsuhiko Hayashi, Yusuke Sakai and Hidetaka Kamigaito

Raga Space Visualization: Analyzing Melodic Structures in Carnatic and Hindustani Music 57
Soham Korade, Suswara Pochampally and Saroja Tk

Musical Ethnocentrism in Large Language Models 62
Anna Kruspe

Analyzing Byte-Pair Encoding on Monophonic and Polyphonic Symbolic Music: A Focus on Musical Phrase Segmentation 69
Dinh-Viet-Toan Le, Louis Bigo and Mikaela Keller

Lyrics for Success: Embedding Features for Song Popularity Prediction 75
Giulio Prevedello, Ines Blin, Bernardo Monechi and Enrico Ubaldi

The Role of Large Language Models in Musicology: Are We Ready to Trust the Machines? . . . 81
Pedro Ramoneda, Emila Parada-Cabaleiro, Benno Weck and Xavier Serra

"Does it Chug?" Towards a Data-Driven Understanding of Guitar Tone Description 87
Pratik Sutar, Jason Naradowsky and Yusuke Miyao

Evaluation of Pretrained Language Models on Music Understanding 98
Yannis Vasilakis, Rachel Bittner and Johan Pauwels

FUTGA: Towards Fine-grained Music Understanding through Temporally-enhanced Generative Augmentation 107
Junda Wu, Zachary Novack, Amit Namburi, Jiaheng Dai, Hao-Wen Dong, Zhouhang Xie, Carol Chen

and Julian McAuley

The Interpretation Gap in Text-to-Music Generation Models 112
Yongyi Zang and Yixiao Zhang

Genre-conformity in the topics of lyrics and song popularity

Anna Aljanaki
University of Tartu
aljanaki@gmail.com

Abstract

The genre of a song defines both musical (rhythmic, timbral, performative) aspects of a song, but also the themes of lyrics and the style of writing. The audience has certain expectations as to emotional and thematic content of the genre they listen to. In this paper we use Music4All database to investigate whether breaking these expectations influences song popularity. We use topic modeling to divide song lyrics into 36 clusters, and apply tag clustering to separate the songs into 15 musical genres. We observe that in some genres (metal, hip-hop) lyrics are mostly written in specific topics, whereas in other genres they are spread over most topics. In most genres, songs that have lyrics that are not representative of the genre, are more popular than songs with lyrics that are more typical for the genre.

1 Introduction

For different listeners, different aspects of a song might be important: rhythm, lyrics, timbre, or epoch that a recording comes from (Huang et al., 2023). Moreover, these principles might vary for different genres and even countries (Schedl et al., 2020). Lyrics seem to be a universally important aspect of a song for many listeners, influencing perceived emotion, enhancing experience and shaping preferences (Alinka Greasley and Sloboda, 2013).

However, lyrics are an often overlooked aspect in music information retrieval. In a recent survey on content-driven music recommendation, 70% of the studies used audio signal in content based music recommender system, while only 30% used any embedded metadata, including lyrics (Deldjoo et al., 2024).

The genre of a song plays an important role in shaping the content of its lyrics. For instance, pop music typically focuses on the topics of romantic love and heartbreak. In hip-hop and rap genres, it is typical for the lyrics to contain slang and obscene vocabulary, and cover topics of social justice and politics.

In (Tsaptsinos, 2017) the songs were classified by genres using lyrics with high accuracy, showing that specific words can be indicative of the genre with a high certainty.

In this paper we will apply topic modeling to song lyrics, and answer the following research questions:

1. How specific are lyrics topics in various genres?

2. Does genre-conformity of the lyrics influence song popularity, and how?

2 Related Work

Lyrics, including topic modeling approaches, have been successfully used for music recommendation (Vystrcilová and Peka, 2020; Patra et al., 2017; Jang et al., 2019; Sasaki et al., 2014). Song popularity prediction was attempted both from audio content (Lee and Lee, 2018) and from lyrics (Martín-Gutiérrez et al., 2020). In (Agatha et al., 2024), lyrics emotional content was estimated using Sentence-BERT transformer, and song popularity was predicted based on that.

3 Data

In this research we will use the lyrics and listening history from music4all database (Santana et al., 2020). This database contains 109269 songs, out of which 91% contain lyrics, and the rest are instrumental. The songs come with more than 5 million listening events, generated by 14127 users.

Out of the songs that have lyrics, 84% are in English, and the rest of the songs are in 44 different languages. For our purposes we do not need instrumental songs and their listening history, hence we discard them. We translate the lyrics in foreign languages into English using Google Cloud Translation API[1]. Some of the English songs contained mixed English and Korean text, and occasional Korean words were translated as well. The cleaned dataset is available at the project repository:[2].

[1] cloud.google.com/translate
[2] github.com/aljanaki/lyrics_topic_analysis

Genre	Amount of songs	Representation kurtosis	Non-representative topics / representative topics	P-value
Rock	22427	-0.99	0.94	0.11
RnB	5140	1.98	0.99	0.96
Punk	3849	-0.6	0.93	0.39
Pop	19555	2.17	**0.75**	**0.01**
New age	1389	14.08	**1.26**	**0.05**
Jazz	959	2.4	**1.45**	**0.05**
Industrial	6033	1.00	1.16	0.06
Hip-hop	4129	27.27	**0.73**	**0.00**
Hardcore	4332	0.18	**1.39**	**0.00**
Funk	5610	11.27	1.08	0.46
Folk	7726	-0.76	1.06	0.38
Electronic	5834	2.65	1.13	0.42
Death metal	6802	8.84	**1.34**	**0.00**
Blues	1089	4.69	1.06	0.75
Black metal	1019	14.34	**1.28**	**0.00**

Table 1: Statistics per genre. Amount of songs: how many songs in the dataset were in that genre. Representation kurtosis: kurtosis computed on song amounts in lyrics topics. Non-representative/representative topics: average play counts ratio in less-representative for this genre divided by average play counts in topics more representative for this genre. P-value: result of the t-test for play counts comparison between these groups.

3.1 Defining genres

In Music4all database, genre labels have a very large cardinality (there are 853 different genres). Most of these genres are represented only by a few songs, whereas each song is usually annotated with several genres, which were scraped from a website by the dataset creators. In order to make analysis by genre possible, we clustered these fine-grained genres using the following process:

1. We computed genre co-occurrence matrix C on the song by genre matrix.
2. Based on matrix C, we computed genre pairwise cosine similarity matrix S.
3. Next, we applied hierarchical clustering of the rows (genres) of S.

In this way, we were able to reduce 853 sparsely used genres to just 15 genre clusters. Clusters were labeled manually by selecting the most frequent parent genre in the cluster. E.g., a cluster containing 'avant-garde black metal', 'greek black metal' and 'usbm', along with 23 other similar tags, was named 'black metal'. Unfortunately, two of the 15 clusters, which we named *new age* and *industrial*, were rather eclectic. 21% of songs in the dataset fell into rock cluster, next by popularity were pop, electronic and folk. The smallest genres are blues, black metal and jazz, which had a little over 1000 songs each.

3.2 Lyrics topic modeling

We applied topic modeling to lyrics of songs using Bertopic[3] approach:

1. We computed sentence embeddings using a MPNet sentence transformer from HuggingFace[4]. The embeddings were computed on a complete lyrics text, treating it as a single paragraph of text, creating 768-dimensional embedding vector that represented each song.
2. We applied dimensionality reduction on these embeddings with UMAP, extracting 50 components using cosine similarity.
3. The dimensionally reduced embeddings were clustered using K-means with $k = 40$.

We also experimented with HDBSCAN and BIRCH clustering algorithms, but they did not result in satisfactory clusters. K-means was able to create clusters of roughly equal size and not as many outliers as HDBSCAN.

We inspected the topics in various ways (extracting influential words with TF-IDF, BERT, and

[3] https://maartengr.github.io/BERTopic/index.html

[4] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

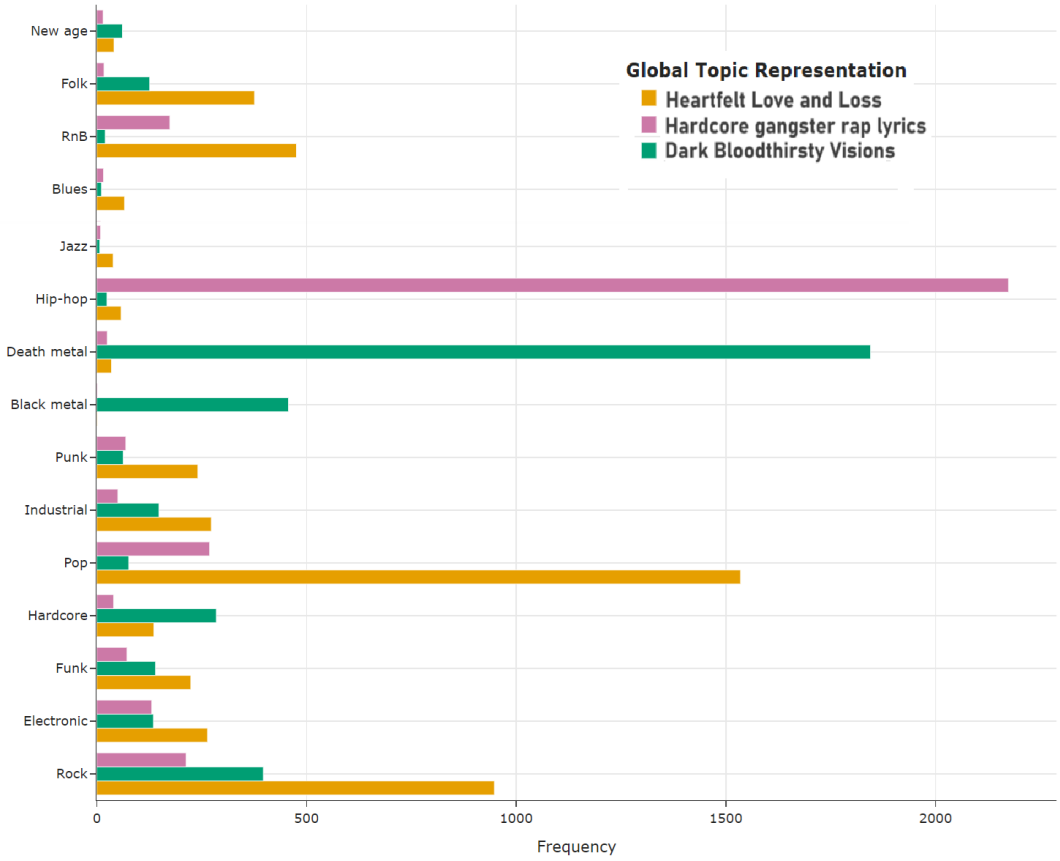


Figure 1: Distribution of three selected lyrics topics over all the genres.

using openAI to label the topics (see Figure 1)). The most popular topics were related to themes of love, unrequited love and breakup. There were also clearly separable topics with obscene lyrics, and lyrics with dark epic themes on death, war and adventure. Four topics were very small, containing less than 50 songs, and were removed, leaving 36 topics.

4 Results

In this section we describe the results that were computed on the following processed data: for each song, we determined a genre cluster that the song belongs to, a lyrics topic, and aggregated play counts for that song over a whole period reflected in the dataset.

4.1 Genre lyrics specificity

For each genre, we will compute how specific are the topics described by the lyrics of the songs to this genre. In order to do that, for each topic c we compute which percentage of the songs from genre g belong to this topic:

$$genre_t = \frac{|genre_g \cap topic_c|}{|topic_c|} \quad (1)$$

In such a way, for each genre, we obtain a vector of values. If lyrics from that genre only belong to a few topics, we will observe large kurtosis of this vector (such as for Hip-Hop or Funk). If the lyrics are spread uniformly across various topics, we will observe small kurtosis (such as for large diverse genres like Pop and Rock).

From Table 1 we can see that such genres as *death metal*, *funk* or *Hip-Hop* have large kurtosis and *rock*, *pop* or *industrial*, which contain a lot of songs and sub-genres, and therefore can either be difficult to define, or were eclectic to begin with, have small kurtosis and are spread over most topics.

4.2 Song popularity vs song lyrics genre conformity

Next, for each genre, we will compute whether conforming to the usual topics of this genre is beneficial for song popularity (increased play counts). The median play counts per topic vary between

8 for the least popular topic ("Inner Demons and Struggles") to 18 for the most popular topic ("Hardcore gangster rap lyrics"), with significant difference between topic play counts on a Kruskal-Wallis H test ($\chi^2(36) = 1902.45$, $p < 0.0001$).

In most genres, there is a long-tail distribution of non-representative topics, and just one or two most popular topics for each genre. We divide the topics into representative and non-representative topics in such a way, that both representative (containing more songs in that genre) and non-representative have not more than 10% difference in amount of songs. This boundary between representative and non-representative topics is different for each genre. For instance, for *black metal*, there is just one representative topic (**Dark bloodthirsty visions**) which contains more songs than all the rest of the topics combined. However, as we can see from Table 1, most often the most represented topics of a genre do not generate biggest play counts. For most genres, songs in topics that are less usual for this genre, receive more attention from the listeners and are listened to more often. For instance, for *hardcore* genre, the most common topic found in lyrics are **Lone wolf** and **Dark self-discovery**. The most listened songs in that genre are on the topics of **Mid-life crisis** and **Unhappy love**. There are some exceptions to this trend: for *Hip-Hop* and *Pop* songs, the most widespread topics are the most popular. For *Pop*, these are **Desire and longing**, **Reminiscing on past love**, and **Relationships and love**. For *Hip-Hop*, these are **Hardcore gangster rap lyrics**,**Mid-life crisis** and **Pep talk**. In some genres, there is no statistically significant difference in popularity between songs with genre-conforming lyrics and other songs.

5 Conclusion

In this paper we showed that artists writing lyrics vary in how much they restrict themselves to certain topics, depending on musical genre, with the most restricting genres being hip-hop, black metal and new age. Also, we showed that when the lyrics are written in a topic not representative for that genre it has a beneficial effect on popularity, for most genres, but not for hip-hop and rock.

References

Hana Agatha, Farica Putri, and Alethea Suryadibrata. 2024. Sentiment analysis on song lyrics for song popularity prediction using bert. *Ultimatics : Jurnal Teknik Informatika*, 15(2):99–105.

Alexandra Lamont Alinka Greasley and John Sloboda. 2013. Exploring musical preferences: An in-depth qualitative study of adults' liking for music in their personal collections. *Qualitative Research in Psychology*, 10(4):402–427.

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

Shuhua Huang, Chenhao Hu, Weiyang Kong, and Yubao Liu. 2023. Disentangled contrastive learning for knowledge-aware recommender system. In *The Semantic Web – ISWC 2023*, pages 140–158, Cham. Springer Nature Switzerland.

Sein Jang, Battulga Lkhagvadorj, and Aziz Nasridinov. 2019. Preference-aware music recommendation using song lyrics. In *Big Data Applications and Services 2017*, pages 183–195, Singapore. Springer Singapore.

Junghyuk Lee and Jong-Seok Lee. 2018. Music popularity: Metrics, characteristics, and audio-based prediction. *IEEE Transactions on Multimedia*, 20(11):3173–3182.

David Martín-Gutiérrez, Gustavo Hernández Peñaloza, Alberto Belmonte-Hernández, and Federico Álvarez García. 2020. A multimodal end-to-end deep learning architecture for music popularity prediction. *IEEE Access*, 8:39361–39374.

Braja Gopal Patra, Dipankar Das, and Sivaji Bandyopadhyay. 2017. Retrieving similar lyrics for music recommendation system. In *ICON*.

Igor André Pegoraro Santana, Fabio Pinhelli, Juliano Donini, Leonardo Gabiato Catharin, Rafael B. Mangolin, Yandre M. G. Costa, Valéria Delisandra Feltrim, and Marcos Aurélio Domingues. 2020. Music4all: A new music database and its applications. *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 399–404.

Shoto Sasaki, Kazuyoshi Yoshii, Tomoyasu Nakano, Masataka Goto, and Shigeo Morishima. 2014. Lyricsradar: A lyrics retrieval system based on latent topics of lyrics. In *ISMIR*.

Markus Schedl, Christine Bauer, Wolfgang Reisinger, Dominik Kowald, and E. Lex. 2020. Listener modeling and context-aware music recommendation based on country archetypes. *Frontiers in Artificial Intelligence*, 3.

Alexandros Tsaptsinos. 2017. Lyrics-based music genre classification using a hierarchical attention network. *ArXiv*, abs/1707.04678.

Michaela Vystrcilová and Ladislav Peka. 2020. Lyrics or audio for music recommendation? *Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics*.

PIAST: A Multimodal Piano Dataset with Audio, Symbolic and Text

Hayeon Bang[1] Eunjin Choi[1] Megan Finch[1] Seungheon Doh[1]
Seolhee Lee[2] Gyeong-Hoon Lee[2] Juhan Nam[1]

[1]Graduate School of Culture Technology, KAIST, South Korea
[2]NCSOFT, South Korea
{hayeonbang, jech, meganelisabethfinch, seungheondoh, juhan.nam}@kaist.ac.kr, {seolhee, ghlee0304}@ncsoft.com

Abstract

While piano music has become a significant area of study in Music Information Retrieval (MIR), there is a notable lack of datasets for piano solo music with text labels. To address this gap, we present PIAST (PIano dataset with Audio, Symbolic, and Text), a piano music dataset. Utilizing a piano-specific taxonomy of semantic tags, we collected 9,673 tracks from YouTube and added human annotations for 2,023 tracks by music experts, resulting in two subsets: PIAST-YT and PIAST-AT. Both include audio, text, tag annotations, and transcribed MIDI utilizing state-of-the-art piano transcription and beat tracking models. Among many possible tasks with the multimodal dataset, we conduct music tagging and retrieval using both audio and MIDI data and report baseline performances to demonstrate its potential as a valuable resource for MIR research.

1 Introduction

Piano music presents unique opportunities for music research due to its ability to express diverse styles using a single instrument and its superior transcription performance. Given these characteristics, it has become a significant area of study in Music Information Retrieval (MIR), encompassing tasks such as classification (Hung et al., 2021; Chou et al., 2021), and music generation with various conditions (Wu and Yang, 2023; Choi and Lee, 2023). While these tasks require datasets that combine piano audio with various modalities such as MIDI, sheet music, or text, there is a notable scarcity of such comprehensive multimodal piano datasets.

However, existing multimodal music datasets, particularly music-text datasets, rarely focus exclusively on piano music, and piano solo pieces comprise only a small portion of general music-text datasets. For instance, in the ECALS Dataset (Doh et al., 2023), a subset of the Million Song Dataset (Bertin-Mahieux et al., 2011), the number of piano solo tracks is very limited. We observed that excluding tracks tagged with instruments other than the piano or genres that could not be solely represented by the piano, only approximately 0.46% of the entire dataset can be identified as piano solo music.

Several piano datasets, such as MAESTRO (Hawthorne et al., 2019), have been developed in recent years, which provide classical piano performances primarily used for piano transcription. Another classical piano dataset, GiantMIDI (Kong et al., 2022), is also commonly used in transcription tasks. Other datasets like Pop1K7 (Hsiao et al., 2021) focus on the performance generation of pop piano music, while PiJAMA (Edwards et al., 2023) is employed for performer identification tasks with their jazz piano data. However, these datasets are confined to a single genre and lack text labels. This absence of genre diversity within a single dataset and the lack of textual information underscores the need for a piano dataset with text information.

Some piano datasets contain emotion labels, such as EMOPIA (Hung et al., 2021) and VGMIDI (Ferreira and Whitehead, 2019). However, these datasets are annotated only with emotion information based on either Russell's four quadrants (Hung et al., 2021) or the valence-arousal model (Ferreira and Whitehead, 2019). This limited annotation approach lacks the rich textual descriptions needed for text-based MIR tasks.

To address the limitations, we present multimodal piano music data with rich text annotations and transcribed MIDI. To build the dataset, we first created a piano-specific taxonomy with 31 tags that include genre, emotion, mood, and style information to encompass the broad and diverse musical range that the piano can express. Based on this taxonomy, we collected data from YouTube, transcribed it to MIDI format, and conducted an

annotation process.

The PIAST dataset consists of two subsets: **PIAST-YT**, 9,673 tracks collected from YouTube, providing audio and text information (titles, tags, and descriptions), and **PIAST-AT**, 2,023 tracks with annotations by music experts. This dual approach ensures both breadth and accuracy in the dataset. Additionally, PIAST includes transcribed performance MIDI data alongside audio and text, enhancing its capabilities beyond existing methods (Hsiao et al., 2021).

This paper details the dataset collection process and analyzes the data. We present baseline results for piano music annotation and retrieval tasks, utilizing the PIAST-YT and the PIAST-AT datasets across audio and MIDI domains. The PIAST dataset is available in our online repository[1], and the source code for the experiments can be found on GitHub[2].

2 Dataset

2.1 Taxonomy for Piano Music

To encompass and precisely define the range of expressions possible in solo piano music, we constructed a comprehensive taxonomy considering genre, emotion, mood, and style tags. We classified genres suitable for solo piano music into four categories: jazz, classical, new-age, and pop piano covers, defining sub-genres within each. The detailed classification of the classical genre was not included in this dataset due to the extensive range and complexity unique to classical music. For emotion and mood taxonomy, we combined vocabularies from four existing music datasets with emotion tags (Turnbull et al., 2007; Rouhi et al., 2019; Aljanaki et al., 2017; Choi et al., 2022)), eliminating overlaps. Seven music experts who majored in music composition rated the tags on a 1-5 Likert scale for their suitability in describing solo piano music. We included only words scoring 3.5 or higher and established a taxonomy of 39 tags. After the first annotation process, we removed tags with excessively high co-occurrence or low selection frequencies, resulting in a final specialized taxonomy of 31 words for piano music.

2.2 The PIAST-YT Dataset

The PIAST-YT dataset comprises approximately 9,673 tracks (1,006 hours) of audio collected from YouTube, accompanied by text information (title, tags, and descriptions of the video). We employed two collection methods: tag-based and channel-based. The tag-based method used our taxonomy to gather diverse styles of piano music from YouTube. However, the inherent variability in the availability of solo piano content on the platform led to some imbalance in the collected data. To ensure the inclusion of high-quality solo piano content, we also employed a channel-based method, collecting piano performance videos from 23 selected channels known for their piano content. Finally, The PIAST-YT dataset comprises three main components after pre-processing step: audio extracted from videos, text data (titles, tags, and descriptions), and MIDI data generated through transcription.

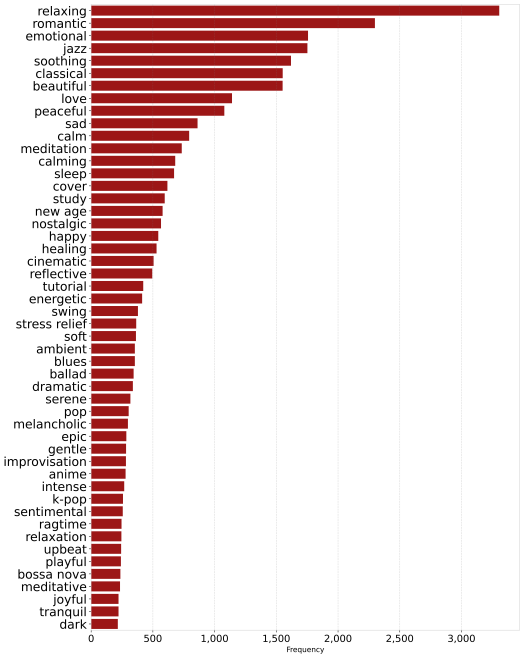


Figure 1: Top 50 words most frequently appearing in the text dataset of the PIAST-YT.

2.2.1 Pre-processing

Audio: To isolate pure piano solo performances, we filtered the data using musicnn (Pons and Serra, 2019), excluding tracks with non-piano sounds in their top 5 tags. Files exceeding 2 hours were removed, and those exceeding 30 minutes were segmented into 10-minute chunks for data consistency. This process reduced the original 1,789 hours of data, about 44%, to 1,006 hours.

Text: The collected text data from YouTube contained diverse and irrelevant information. To extract relevant music-descriptive features, we employed an LLM-based model, specifically ChatGPT 4-Turbo (Ouyang et al., 2022), chosen for its

[1] https://hayeonbang.github.io/PIAST_dataset/
[2] https://github.com/Hayeonbang/PIAST

high performance. This model generated a tag list for each video based on its corresponding text. Figure 1 illustrates the distribution of these generated tags. The total number of vocabulary is 3,160.

MIDI: The piano audio files were transcribed to performance MIDI using an automatic piano transcription model (Kong et al., 2021). The MIDI was then synchronised to beat estimates, and melody and chords were extracted using the Pop1k7 dataset pipeline (Hsiao et al., 2021). For the beat estimates, following (Holzapfel et al., 2012), we used the Mean Mutual Agreement (MMA) between a 'committee' of several state-of-the-art beat trackers, including All-in-One (Kim and Nam, 2023) and madmom (Böck et al., 2014), to filter out samples for which the beat tracking quality was poor. This transcription process was applied to the audio in both the PIAST-YT and the PIAST-AT datasets.

2.3 The PIAST-AT Dataset

Even after processing, the text data in the PIAST-YT exhibited several limitations. Although it was processed using an LLM-based model, it still showed a low correlation with the music content, and some audio files lacked corresponding text data. To address these issues, we created the PIAST-AT, a dataset consisting of piano-specific human-annotated text.

2.3.1 Annotation Process

We stratified 2,400 samples from audio data of PIAST-YT based on the queries used during the collection process, and extracted a 30-second segment from each sample for human annotation. The process involved 15 music experts (7 jazz and 8 classical musicians) with majors in composition. Each segment was assigned to three annotators using a web-based system. The annotators were divided into five groups tagged with 230 segments. Detailed descriptions and examples were provided to all annotators for each tag to ensure consistency. They were also instructed to exclude samples that did not strictly adhere to solo piano criteria or had subtle mood changes. After two rounds of annotation, we collected tags for 2,023 samples (approximately 17 hours of original audio), with 377 samples excluded through this process.

2.3.2 Dataset Analysis & Tag Consensus

Figure 2 shows the distribution of tags in the PIAST-AT dataset, categorized into Mood/Emotion, Genre, and Style. The Style category includes tags associated with performance difficulty and tempo-related mood. Due to the inherent imbalance of collected audio, there is also a disparity in the frequency of sub-genre tags.

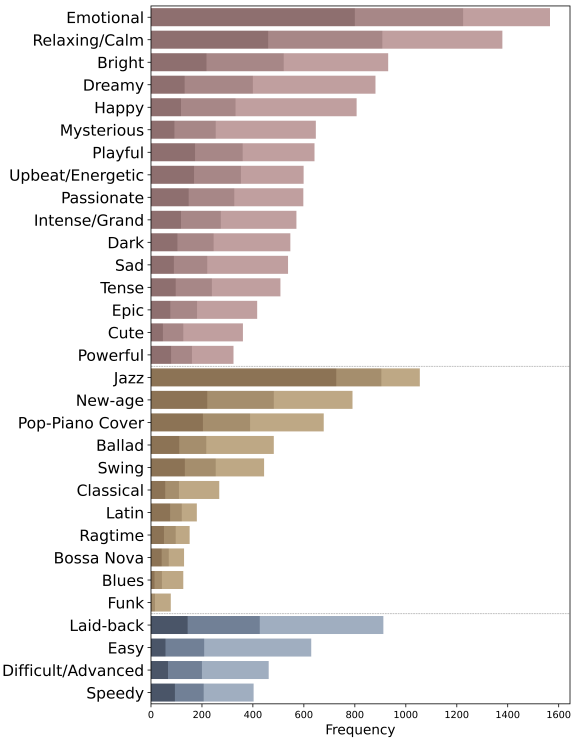


Figure 2: Tag distribution of the PIAST-AT dataset. Three distinct represent the degree of consensus. (Darkest: n=3, Medium: n=2, Lightest: n=1)

The dataset contains the consensus degree among the annotators. To leverage this information, we generated hierarchical captions based on the level of agreement as follows:

> *"This is definitely Jazz genre; (3 agreements)*
> *also Speedy style; also Playful mood; (2 agreements)*
> *potentially Latin genre; potentially Easy style; potentially Bright, Happy, Cute mood of piano music. (1 agreement)"*

The PIAST-AT dataset comprises audio, transcribed MIDI, and text annotations (tags and captions), offering a rich representation of musical characteristics and annotator consensus.

3 Piano Music Classification

In this section, we present the application of our proposed dataset for piano music annotation and retrieval tasks in both the audio and MIDI domains. We employed a two-stage framework: 1) pre-training and 2) transfer learning. For pre-training, we used the PIAST-YT dataset to train

a general-purpose piano-specific model with large-scale audio, MIDI, and diverse text data. We leveraged text supervision through music-text joint embedding pre-training (Huang et al., 2022; Manco et al., 2022; Doh et al., 2024b). In the transfer learning stage, we utilized the PIAST-AT dataset to train a piano classification model as a downstream task.

3.1 Pre-training and Transfer Learning

To develop a piano-specific pre-trained model, we extracted embeddings from audio, MIDI, and text modality encoders. We applied contrastive loss to maximize similarity between corresponding pairs (audio-text or MIDI-text) while minimizing similarity with in-batch negative samples. Following previous studies (Huang et al., 2022; Manco et al., 2022; Doh et al., 2024b), each encoder consists of a modality-specific backbone, a linear projection layer, and an l_2 normalization layer. We used a modified ResNet-50 (Radford et al., 2021) for audio, RoBERTa (Liu et al., 2019) for text, and MidiBERT-Piano (Chou et al., 2021) with average pooling for MIDI.

For the classification model, we employed the probing protocol (Doh et al., 2023; Castellon et al., 2021). We used the pre-trained audio and MIDI encoders as frozen feature extractors and trained linear models and one-layer MLPs as shallow classifiers on top of them, with 512 hidden states and ReLU activation.

3.2 Implementation Details

We processed the input data for pre-training and transfer learning as follows: Audio inputs were 10-second signals sampled at 22050 Hz, converted to log-mel spectrograms with 128 mel bins using a 1024-point FFT with a Hann window and a 10 ms hop size. For MIDI, pre-processed MIDI files were converted to the CP (*compound word*) representation (Hsiao et al., 2021) and fed into a 12-layer BERT with a maximum sequence length of 512. All models were optimized using AdamW with a 5e-5 learning rate, and a dropout rate of 0.4 applied to the audio classification model. We used batch sizes of 128 for audio and 48 for MIDI data during pre-training. Pre-training models were trained for 150 epochs, while classification models ran for 700 epochs, with the best model selected based on validation loss. The PIAST-AT dataset was split into 80% for training, 10% for validation, and 10% for testing sets.

	Music → Tag		**Tag → Music**	
	ROC-AUC	**PR-AUC**	**ROC-AUC**	**PR-AUC**
Supervised				
Audio	81.06	70.71	73.22	50.97
MIDI	84.82	75.24	79.14	58.00
Pre-train and Transfer Learning				
Audio	84.52	74.73	79.01	58.70
MIDI	85.69	76.27	80.63	61.53

Table 1: Performance results for music-to-tag and tag-to-music tasks.

3.3 Evaluation & Results

We evaluated our classification models on two tasks: the *annotation task*, which involves finding appropriate tags for given music, and the *retrieval task*, which focuses on finding suitable music for provided tags. Following previous studies (Choi et al., 2019; Doh et al., 2024a), we employed the area under the ROC and PR curves averaged over instances as evaluation metrics for the annotation task. The retrieval task was assessed using the area under the ROC and PR curves averaged over labels. To demonstrate the effectiveness of the proposed large PIAST-YT dataset, we used a supervised model trained exclusively on the smaller PIAST-AT dataset as the baseline model.

Table 1 compares the annotation and retrieval performance across 1) audio and MIDI modalities, and 2) the supervised versus pre-train and transfer framework. The MIDI model consistently outperformed the audio model across both tasks. Pre-training with PIAST-YT improved the performance of both models on all metrics, demonstrating its effectiveness. This pre-training approach led to superior performance in both music-to-tag and tag-to-music tasks.

4 Conclusion

In this paper, we introduced PIAST, a piano dataset with audio, symbolic and text. Our experiments demonstrated the dataset's effectiveness for piano music annotation and retrieval tasks, showing improvements with pre-training. The PIAST dataset supports various applications, including improved music retrieval, text-based music generation, music analysis, and emotion/genre classification. To further enhance the dataset, our future work will address tag imbalances by adding more samples and incorporating additional processed data such as lead sheets and chord annotations.

Acknowledgments

This work was supported by the collaboration with NCSOFT, Korea.

References

Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. 2017. Developing a benchmark for emotional analysis of music. *PloS one*, 12(3):e0173392.

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*.

Sebastian Böck, Florian Krebs, and Gerhard Widmer. 2014. A multi-model approach to beat tracking considering heterogeneous music styles. In *Proceedings of 15th International Conference on Music Information Retrieval (ISMIR)*.

Rodrigo Castellon, Chris Donahue, and Percy Liang. 2021. Codified audio language modeling learns useful representations for music information retrieval. In *Proceedings of the 22th International Society for Music Information Retrieval Conference (ISMIR)*.

Eunjin Choi, Yoonjin Chung, Seolhee Lee, JongIk Jeon, Taegyun Kwon, and Juhan Nam. 2022. YM2413-MDB: A multi-instrumental fm video game music dataset with emotion annotations. In *Proceedings of 23th International Conference on Music Information Retrieval (ISMIR)*.

Jeong Choi, Jongpil Lee, Jiyoung Park, and Juhan Nam. 2019. Zero-shot learning for audio-based music classification and tagging. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*.

Jongho Choi and Kyogu Lee. 2023. Pop2Piano: Pop audio-based piano cover generation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yi-Hui Chou, I Chen, Chin-Jui Chang, Joann Ching, Yi-Hsuan Yang, et al. 2021. MidiBERT-Piano: Large-scale pre-training for symbolic music understanding. *arXiv preprint arXiv:2107.05223*.

SeungHeon Doh, Jongpil Lee, Dasaem Jeong, and Juhan Nam. 2024a. Musical word embedding for music tagging and retrieval. *arXiv preprint arXiv:2404.13569*.

SeungHeon Doh, Minhee Lee, Dasaem Jeong, and Juhan Nam. 2024b. Enriching music descriptions with a finetuned-llm and metadata for text-to-music retrieval. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 826–830. IEEE.

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

Drew Edwards, Simon Dixon, and Emmanouil Benetos. 2023. Pijama: Piano jazz with automatic midi annotations. *Transactions of the International Society for Music Information Retrieval*.

Lucas N Ferreira and Jim Whitehead. 2019. Learning to generate music with sentiment. In *Proceedings of 20th International Conference on Music Information Retrieval (ISMIR)*.

Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. 2019. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *Proceedings of 7th International Conference on Learning Representations (ICLR)*.

Andre Holzapfel, Matthew EP Davies, José R Zapata, João Lobato Oliveira, and Fabien Gouyon. 2012. Selective sampling for beat tracking evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(9):2539–2548.

Wen-Yi Hsiao, Jen-Yu Liu, Yin-Cheng Yeh, and Yi-Hsuan Yang. 2021. Compound Word Transformer: Learning to compose full-song music over dynamic directed hypergraphs. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, volume 35, pages 178–186.

Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. 2022. MuLan: A joint embedding of music audio and natural language. In *Proceedings of 23th International Conference on Music Information Retrieval (ISMIR)*.

Hsiao-Tzu Hung, Joann Ching, Seungheon Doh, Nabin Kim, Juhan Nam, and Yi-Hsuan Yang. 2021. EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. In *Proceedings of 22th International Conference on Music Information Retrieval (ISMIR)*.

Taejun Kim and Juhan Nam. 2023. All-in-one metrical and functional structure analysis with neighborhood attentions on demixed audio. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5. IEEE.

Qiuqiang Kong, Bochen Li, Jitong Chen, and Yuxuan Wang. 2022. GiantMIDI-Piano: A large-scale midi dataset for classical piano music. *Transactions of the International Society for Music Information Retrieval*, 5(1):87–98.

Qiuqiang Kong, Bochen Li, Xuchen Song, Yuan Wan, and Yuxuan Wang. 2021. High-resolution piano transcription with pedals by regressing onset and offset times. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3707–3717.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. 2022. Contrastive audio-language learning for music. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Jordi Pons and Xavier Serra. 2019. musicnn: Pretrained convolutional neural networks for music audio tagging. In *Late Breaking/Demo in the 20th International Society for Music Information Retrieval (ISMIR)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763.

Amirreza Rouhi, Micol Spitale, Fabio Catania, Giulia Cosentino, Mirko Gelsomini, and Franca Garzotto. 2019. Emotify: emotional game for children with autism spectrum disorder based-on machine learning. In *Companion Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 31–32.

Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. 2007. Towards musical query-by-semantic-description using the cal500 data set. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 439–446.

Shih-Lun Wu and Yi-Hsuan Yang. 2023. Compose & Embellish: Well-structured piano performance generation via a two-stage approach. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Lyrics Transcription in Western Classical Music with Whisper: A Case Study on Schubert's Winterreise

Hans-Ulrich Berendes and Simon Schwär and Meinard Müller
International Audio Laboratories Erlangen

Abstract

Automatic Lyrics Transcription (ALT) aims to transcribe sung words from music recordings and is closely related to Automatic Speech Recognition (ASR). Although not specifically designed for lyrics transcription, the state-of-the-art ASR model Whisper has recently proven effective for ALT and various related tasks in music information retrieval (MIR). This paper investigates Whisper's performance on Western classical music, using the "Schubert Winterreise Dataset." In particular, we found that the average Word Error Rate (WER) with the unmodified Whisper model is 0.56 for this dataset, while the performance varies greatly across songs and versions. In contrast, spoken versions of the song lyrics, which we recorded, are transcribed with a WER of 0.14. Further systematic experiments with source separation and time-scale modification techniques indicate that Whisper's accuracy in lyrics transcription is less affected by the musical accompaniment and more by the singing style.

1 Introduction

Lyrics, the words of a song, are vital to vocal music. They contain important information for listeners and bridge the gap between music and language. Automatic Lyrics Transcription (ALT) extracts these words, often from a mix of instruments and vocals (Tsai et al., 2018). Automatic Speech Recognition (ASR) performs a similar task for normal speech (Malik et al., 2021). While both involve processing the human voice, speech, and singing differ in pitch fluctuations, pronunciation, speed, time variations, and vocabulary (Humphrey et al., 2019). Musical accompaniment can further complicate ALT, as it superimposes the singing voice, often with high temporal and spectral correlations (Gupta et al., 2020). Due to these differences, ASR and ALT have long been considered separate tasks (Kruspe, 2024).

Recent ASR advances rely on large, diverse datasets and often use weakly-supervised or self-supervised training (Baevski et al., 2020; Peng et al., 2024). One state-of-the-art model, Whisper, is trained on a total of 5 million hours of data (Radford et al., 2023). Trained on such extensive data, Whisper shows promising capabilities for ALT as well. It can either be used without modifications (Cífka et al., 2023), in combination with a Large Language Model (LLM) for transcript post-processing (Zhuo et al., 2023) or be fine-tuned on specific music genres (Wang et al., 2023). Understanding large pre-trained models is crucial, as these models can be useful for tasks with limited data like ALT, in particular for underrepresented languages or genres (Latif et al., 2023; Wang et al., 2024).

This paper aims to better understand Whisper's ALT performance and the challenges of transcribing singing compared to speech. Different from the other works mentioned above, we focus on Western classical music. In particular, we use the "Schubert Winterreise Dataset" (SWD) (Weiß et al., 2021) as a case study, which comprises nine complete recordings of the 24-song cycle "Winterreise" by Franz Schubert. The Winterreise is composed for solo voice with piano accompaniment, based on German poems from the early 19$^{\text{th}}$ century.

Our contributions are twofold: an in-detail analysis of Whisper's ALT performance on the SWD, and a comparison of speech and singing transcription through experiments with spoken versions of the lyrics, source separation, and time-scale modification.

2 Experimental Setup

2.1 Whisper

The multilingual ASR model Whisper, introduced by Radford et al. (2023), is based on a transformer architecture and available in various sizes. In this

work, we use the largest and latest pre-trained version, `large-v3` [1]. For simplicity, we refer to this model as Whisper. It has been trained on 4 million hours of unlabeled data and 1 million hours of weakly-supervised data, both not publicly available. Despite being tailored for ASR, there are indications that music is included to some extent in the training data (Zhuo et al., 2023). Although there has been work on improving Whisper for ALT (Zhuo et al., 2023; Wang et al., 2023, 2024), we use the model in its original state to better understand its behavior and potentially evaluate differences between speech and singing.

2.2 Evaluation Dataset

The SWD (Weiß et al., 2021), contains nine commercial recordings of all 24 songs of the Winterreise. These versions feature different male singers, pianos, acoustic conditions, and audio quality. The total number of words per version in the lyrics is 2644. In the following, we denote the songs using their respective number ranging from `SWD-01` to `SWD-24`. Following the dataset paper, we denote the versions with a two-letter identifier alongside the recording year, e.g., `AL98`. For more details on the versions, see Weiß et al. (2021). Since Whisper's training data is not public, we cannot ensure that there is no overlap with the publicly available SWD. We use this dataset because we consider the classical singing style together with the accompaniment to be a challenging scenario for an ASR system. Additionally, the SWD enables cross-version analysis by offering multiple performances of each piece.

2.3 Evaluation Metrics

The most commonly used metric to measure the accuracy of ASR and ALT is the Word Error Rate (WER) (Malik et al., 2021). Given a reference text and a transcript, it is defined as

$$\mathrm{WER} = \frac{D + I + S}{R}, \tag{1}$$

where D is the number of deletions, I the number of insertions, S the number of substitutions, and R the number of words in the reference text. The WER can exceed 1 when a transcript has more words than its reference. While our focus lies on the WER, we additionally report the Character Error Rate (CER) for a more fine-grained analysis. It is defined similarly to the WER but on a character level, rather than a word level. To ensure consistency, we standardize both the reference and transcript texts by removing all punctuation and capitalization before calculating the metrics. Considering the stochastic decoding in the Whisper model, we average the metrics over five independent trials to ensure result stability, as done in Cífka et al. (2023). We will briefly discuss the impacts of this in Section 3.

[1] Available at https://github.com/openai/whisper/

3 Lyrics Transcription Results

In this section, we evaluate the transcription performance of Whisper for singing with accompaniment. Figure 1 shows the WER of the Whisper transcription of SWD for each song and each version, along with the respective averages. The overall mean WER is $\mu = 0.56$ but we can see considerable differences, both across songs and versions with an overall standard deviation of $\sigma = 0.234$.

3.1 Results across Versions

We first investigate the differences between versions. The average WER varies from $\mu_{\texttt{FI66}} = 0.49$ to $\mu_{\texttt{AL98}} = 0.64$, an absolute difference of up to 0.15 for the same songs and lyrics. The standard deviation is $\sigma_{\text{version}} = 0.044$. Notably, the oldest recording `HU33` (with the worst audio quality) has a mean WER of $\mu_{\texttt{HU33}} = 0.54$, just below the average, indicating Whisper's robustness against poor audio quality (Radford et al., 2023). No version consistently gives better or worse results. For example, `FI66` has the lowest average WER but shows the highest WER of 0.46 for `SWD-02` and the lowest WER of 0.24 for `SWD-05`.

3.2 Results across Songs

Next, we examine WER variations across songs. The mean WER (across versions) ranges from $\mu_{\texttt{SWD-02}} = 0.29$ to $\mu_{\texttt{SWD-21}} = 0.98$, an absolute difference of 0.69. The standard deviation of per-song averages is $\sigma_{\text{song}} = 0.148$, larger than $\sigma_{\text{version}} = 0.044$ mentioned above.

For deeper insight, we examine songs `SWD-02` and `SWD-21`. Musically, `SWD-02` features a fast tempo with subtle piano accompaniment, mainly supporting the voice. Figure 2 shows the lyrics of the first two stanzas of `SWD-02` alongside the corresponding transcript. Many errors are substitutions, e.g., "Wetterfahne" becomes "Wetterfalle". Whisper also struggles with compound words, e.g.,

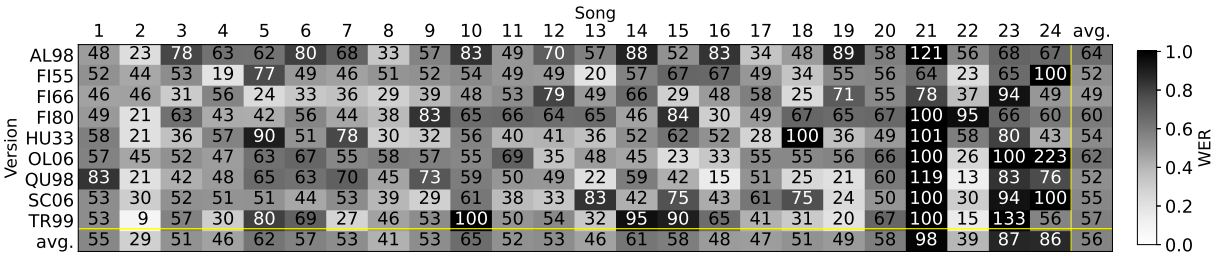


Figure 1: WER of each song and version in SWD, sorted by averages over songs and versions. For better visibility, the numbers are given in 100·WER.

"Liebchens Haus" is written as "Liebchenshaus", and "nimmer" is split into "nie mehr", semantically equivalent in German.

SWD-21 has the highest average WER, with seven out of nine versions showing a WER of 1.0 or higher. In these instances, Whisper often fails to produce meaningful transcriptions. The song has a slow tempo with long piano-only sections. Transcripts frequently contain irrelevant text, such as music descriptors ("Piano Music") or unrelated phrases ("Thank you for listening"), an issue already previously documented (Cífka et al., 2023; Zhuo et al., 2023).

Reference	Transcript
Der Wind spielt mit der Wetterfahne auf meines schönen Liebchens Haus. Da dacht ich schon in meinem Wahne, sie pfiff' den armen Flüchtling aus. Er hätt es eher bemerken sollen, des Hauses aufgestecktes Schild, so hätt er nimmer suchen wollen im Haus ein treues Frauenbild.	Der Wind spielt mit der Wetterfalle auf meine schöne Liebchenshaus. Verdacht ich schon in meinem Wale, sie pfiff den armen Flüchtling aus. Er hätte sicher bemerken sollen, des Hauses aufgestellte Schild. So hätte er nie mehr suchen wollen, im Haus ein treues Frauenbild.

Figure 2: Comparison between the first two stanzas of SWD-02 (reference text on the left) and a Whisper-generated transcript (on the right), with errors highlighted in red. The WER of this excerpt is 0.33.

3.3 Discussion

The average WER of 0.56 on the SWD is considerably higher compared to speech benchmark datasets for long-form transcription, which are in the range of 0.04 to 0.2 (Radford et al., 2023). Cífka et al. (2023) utilized Whisper for ALT with a variety of modern genres, including rock and pop music, and reported a WER of 0.36, which is still significantly lower than our results. This suggests that the music in the SWD presents a more challenging task compared to rock and pop music. Although it is difficult to reason about errors of black box systems like Whisper, we hypothesize that some errors, e.g., seen in Figure 2, can be attributed to the poetic style and old language.

Whisper's stochastic decoding introduces noise, leading to some uncertainty in our results. Averaging over five trials, the average standard deviation is 0.13, with a confidence of 0.06 for the mean WER of a single track. We argue that this is sufficiently small to maintain the validity of our observations.

4 Comparative Analysis of Speech and Singing Transcription

In the previous section, we have seen that the ALT performance of Whisper for the classical music dataset SWD is low compared to speech benchmarks. In this section, we further explore this difference, by investigating possible factors that may deteriorate the transcription performance from speech to classical singing.

4.1 Influence of Musical Accompaniment

One major difference between ASR and ALT is the musical accompaniment, which acts as a correlated "background noise" when transcribing singing. The varying accompaniment could potentially account for WER differences between ALT and ASR datasets, as well as differences between the songs in the SWD. To test this hypothesis, we employ Musical Source Separation (MSS) to extract vocal tracks of the SWD, which we denote with V-MSS. For source separation, we use the commercial system provided by the company *AudioShake*, further denoted by $\texttt{V-MSS}_{\texttt{AS}}$, as well as the open-source model hybrid Demucs introduced in Défossez (2021), denoted by $\texttt{V-MSS}_{\texttt{HDMC}}$. In Table 1 we report the respective WERs and CERs. The MSS pre-processing does not improve the results significantly, which aligns with previous findings by Cífka et al. (2023). This indicates that the musical accompaniment is not the primary source of errors. However, small artifacts introduced by the MSS algorithms could be detrimental to the transcription performance and more work on clean multi-track data could give more insight into this

	MIX	V-MSS$_{AS}$	V-MSS$_{HDMC}$	V-SP
WER [%]	56.1	54.1	55.6	14.6
CER [%]	44.3	42.4	43.9	9.4

Table 1: WER and CER for the three signal types: unprocessed polyphonic input (MIX), vocals extracted with MSS (V-MSS) with further indication of the used MSS system, and the spoken version of the lyrics (V-SP).

aspect. The robustness against musical accompaniment might not hold for models other than Whisper, since previous work has shown that jointly training an MSS system with a lyrics transcriber model can be beneficial (Gao et al., 2023).

4.2 Sung vs. Spoken Lyrics

There is a plethora of work, comparing the acoustic differences between speech and singing, e.g., List (1963); Patel et al. (2006); Gao et al. (2018); Vanden Bosch der Nederlanden et al. (2023). We want to directly compare these two domains in terms of Whisper's respective transcription accuracy. To this end, we recorded spoken versions of the song lyrics for the SWD, which we denote with V-SP. Our recordings feature two speakers, male and female, both native German speakers.

Table 1 shows the WER and CER for the spoken lyrics (V-SP). The WER for V-SP is 0.146 and the CER is 0.094 and therefore considerably lower, compared to the original SWD. Therefore we can rule out the distinct vocabulary of the SWD as the single source of errors. Since MSS pre-processing did not improve the results, we hypothesize that singing itself, and particularly the classical singing style in the SWD, poses a challenge for Whisper. One distinct difference between speech and singing is the duration of individual phonemes (Kruspe, 2024). To investigate the influence of this, we apply time-scale modification to the spoken lyrics V-SP, using the `libtsm` Python package[2], based on Driedger and Müller (2014). Note that this introduces artifacts, which grow more noticeable with stronger modification, however, the pitch is not changed. Each time-scale modified signal is characterized by a single time stretch factor, where a value smaller than 1 denotes a higher speed compared to the original signal. Figure 3 shows the average WER across various time stretch factors.

The transcription performance decreases for very high or low time-stretch factors but remains fairly robust to small changes. This suggests that strong

[2] https://github.com/meinardmueller/libtsm

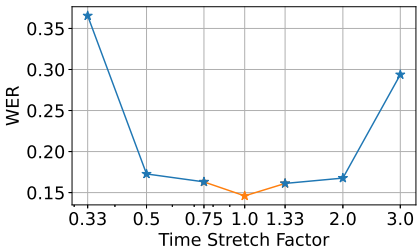


Figure 3: Average WER for time-stretched speech signals across various time stretch factors. The orange color denotes the unaltered speech.

deviations from normal speech are problematic for Whisper, which seems reasonable given its speech-focused training. Quantifying phoneme duration deviations from normal speech depends on the music genre and language, but stretching factors of 3 are common for vowels (Duan et al., 2013; de Medeiros and Cabral, 2018). Our time-stretching experiment may explain why the SWD is challenging for Whisper. Further experiments, analyzing correlations between errors and stretched phonemes could help adapt ASR models to singing.

5 Conclusion and Future Work

Our study investigates Whisper's ALT performance on Western classical music using the SWD as a case study. We find a higher WER for the SWD compared to speech or other singing datasets, with significant fluctuations across songs and versions. Vocabulary has a minor impact, as spoken lyrics WER is comparable to other speech benchmarks. MSS-based vocal extraction has a negligible influence on the WER, indicating musical accompaniment is also not the primary issue. Preliminary experiments show that Whisper is robust against small speed variations but sensitive to larger variations in talking speed compared to normal speech.

We hope our study serves as a starting point for analyzing how characteristics of speech and singing influence ASR model performance. Our evaluation methodology, though applied to Whisper, is relevant beyond a single model. Applying this approach to other ASR models, such as Peng et al. (2024), could enhance understanding of their behavior. This perspective positions our work as a case study for evaluating large audio models and highlights the potential of the music domain for thorough analysis of pre-trained models.

Acknowledgements: This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant No. 328416299 (DFG MU 2686/10-2) and Grant No. 401198673 (DFG MU 2686/13-2). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Ondřej Cífka, Constantinos Dimitriou, Cheng-i Wang, Hendrik Schreiber, Luke Miner, and Fabian-Robert Stöter. 2023. Jam-ALT: A Formatting-Aware Lyrics Transcription Benchmark. *arXiv*, abs/2311.13987.

Beatriz Raposo de Medeiros and Joao Paulo Cabral. 2018. Acoustic distinctions between speech and singing: Is singing acoustically more stable than speech? In *Proceedings of the International Conference on Speech Prosody*, pages 542–546, Poznań, Poland.

Alexandre Défossez. 2021. Hybrid spectrogram and waveform source separation. In *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, pages 1–13, Online.

Jonathan Driedger and Meinard Müller. 2014. TSM Toolbox: MATLAB implementations of time-scale modification algorithms. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pages 249–256, Erlangen, Germany.

Zhiyan Duan, Haotian Fang, Bo Li, Khe Chai Sim, and Ye Wang. 2013. The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2013, Kaohsiung, Taiwan, October 29 - November 1, 2013*, pages 1–9.

Xiaoxue Gao, Chitralekha Gupta, and Haizhou Li. 2023. Polyscriber: Integrated fine-tuning of extractor and lyrics transcriber for polyphonic music. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:1968–1981.

Xiaoxue Gao, Berrak Sisman, Rohan Kumar Das, and Karthika Vijayan. 2018. NUS-HLT Spoken Lyrics and Singing (SLS) Corpus. In *International Conference on Orange Technologies (ICOT)*, pages 1–6, Nusa Dua, Indonesia.

Chitralekha Gupta, Emre Yılmaz, and Haizhou Li. 2020. Automatic lyrics alignment and transcription in polyphonic music: Does background music help? In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 496–500, Barcelona, Spain.

Eric J. Humphrey, Sravana Reddy, Prem Seetharaman, Aparna Kumar, Rachel M. Bittner, Andrew Demetriou, Sankalp Gulati, Andreas Jansson, Tristan Jehan, Bernhard Lehner, Anna Krupse, and Luwei Yang. 2019. An introduction to signal processing for singing-voice analysis: High notes in the effort to automate the understanding of vocals in music. *IEEE Signal Processing Magazine*, 36(1):82–94.

Anna Kruspe. 2024. More than words: Advancements and challenges in speech recognition for singing. In *Conference on Electronic Speech Signal Processing (ESSV) Keynote*, Regensburg, Germany.

Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Heriberto Cuayáhuitl, and Björn W Schuller. 2023. Sparks of large audio models: A survey and outlook. *arXiv*, abs/2308.12792.

George List. 1963. The boundaries of speech and song. *Ethnomusicology*, 7(1):1–16.

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80:9411–9457.

Aniruddh D Patel, John R Iversen, and Jason C Rosenberg. 2006. Comparing the rhythm and melody of speech and music: The case of british english and french. *The Journal of the Acoustical Society of America*, 119(5):3034–3047.

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, et al. 2024. Owsm v3. 1: Better and faster open whisper-style speech models based on e-branchformer. *arXiv*, abs/2401.16658.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 28492–28518.

Che-Ping Tsai, Yi-Lin Tuan, and Lin-Shan Lee. 2018. Transcribing lyrics from commercial song audio: the first step towards singing content processing. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5749–5753.

Christina M Vanden Bosch der Nederlanden, Xin Qi, Sarah Sequeira, Prakhar Seth, Jessica A Grahn, Marc F Joanisse, and Erin E Hannon. 2023. Developmental changes in the categorization of speech and song. *Developmental Science*, 26(5):e13346.

Jun-You Wang, Chon-In Leong, Yu-Chen Lin, Li Su, and Jyh-Shing Roger Jang. 2023. Adapting pre-trained speech model for mandarin lyrics transcription and alignment. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8, Taipei, Taiwan.

Jun-You Wang, Chung-Che Wang, Chon-In Leong, and Jyh-Shing Roger Jang. 2024. Mir-mlpop: A multilingual pop music dataset with time-aligned lyrics and audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1366–1370, Seoul, Korea, Republic of.

Christof Weiß, Frank Zalkow, Vlora Arifi-Müller, Meinard Müller, Hendrik Vincent Koops, Anja Volk, and Harald Grohganz. 2021. Schubert Winterreise dataset: A multimodal scenario for music analysis. *ACM Journal on Computing and Cultural Heritage (JOCCH)*, 14(2):25:1–18.

Le Zhuo, Ruibin Yuan, Jiahao Pan, Yinghao Ma, Yizhi Li, Ge Zhang, Si Liu, Roger B. Dannenberg, Jie Fu, Chenghua Lin, Emmanouil Benetos, Wenhu Chen, Wei Xue, and Yike Guo. 2023. LyricWhiz: Robust multilingual zero-shot lyrics transcription by Whispering to ChatGPT. In *Proceedings of the International Society for Music Information Retrieval Conference, ISMIR*, pages 343–351, Milano, Italy.

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 InstructGPT 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. However, 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 multimodal (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 repurpose 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	**62.8 ± 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	**47.9 ± 0.3**
MC	**89.4 ± 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$_{reco}$	**77.7 ± 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 ± 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.

[2] https://ai.meta.com/blog/meta-llama-3/

[3] Similar results were obtained for `MPNet`.

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.

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 / singer-songwriter* [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`.

NLP Analysis of Environmental Themes in Phish Lyrics Across Concert Locations and Years

Anna Farzindar and Jason Jarvis
Loyola Marymount University
1 LMU Drive Los Angeles, California, 90045
anna.farzindar@lmu.edu , jason.jarvis@lmu.edu

Abstract

This work studies the application of advanced AI and natural language processing (NLP) techniques, to analyze the lyrics of Phish, a renowned American jam band known for their groundbreaking improvisational live shows and eclectic lyrics. Focusing on environmental themes within their extensive repertoire, this paper aims to uncover latent topics pertaining to environmental discourse, by using the topic modeling and environmental classifier to filter out the list of topics present within their songs. Through meticulous preprocessing, modeling, and interpretation, our findings shed light on the multifaceted portrayal of environmental issues in Phish's lyrics. In this study, our primary contribution lies in lyrical analysis, as well as visualization and interpretation of the topics their lyrics cover, over the forty plus years the band has existed. Our lyrical visualizations aim to facilitate an understanding of how Phish selects the timing and location for their live performances in relation to the themes present in their music.

1 Introduction

As the planet plunges headlong into 1.5 degree Celsius warming, public support for protecting the biosphere and altering our relationship with the non-human world is critical. To this end, art and music can play a significant role in raising public awareness. We contend that Phish is an important band leading this charge in the United States. Phish, a jam band renowned for their psychedelic rock performances, has connected with their fans over the years through their distinctive approach to improvisation.

Phish is an iconic act in the "jam band" genre of music that has its roots in bebop, bluegrass and rock. Jam bands have several critical characteristics: (1) they allow live recording and sharing of their concerts by fans, (2) they engage in extensive improvisation during live concerts and (3) concert setlists that vary from show to show. Consequently, a jam band never plays the same show twice.

Phish is one of the most significant bands in America despite having almost no radio airplay or music industry awards over their 41-year history. Nonetheless, in 2023 they were the 10th highest grossing rock music touring act in America: selling 597,000 tickets across 41 shows with a total gross of $76.8 million dollars (Frankenberg, 2023). For context, ninth place went to the Red Hot Chili Peppers while Coldplay and Elton John were the first and second respectively.

One of the things Phish is known for is their regular use of environmental themes. This was evident at Phish's 2024 residency in the Las Vegas Sphere. Over the course of the four night stand, Phish organized their shows around a progression of environmental themes "from solid to liquid to gas to plasma" (Renner Brown, 2024). Only the second band to play the venue after U2, the concerts were a groundbreaking success, famously causing attendee (and host of The Price is Right) Drew Carey to say that Phish's shows made U2 "look like a bar band" and that he "wanted to call U2 and get my money back" (Simpson, 2024).

Also on April 23, 2022 (Earth Day), the band performed what is now a legendary show at Madison Square Garden built around the theme of water that included whale and dolphin drones circling the arena while bathed in blue light that simulated the ocean with kelp descending from the ceiling as the band played. Phish does more than just create images of nature in their songs. Notably, Phish's non-profit organization, The Waterwheel Foundation specifically lists environmental issues as a key area that donations support[1].

In 2023, the band held two benefit concerts in New York raising $3.5 million for people suffering due to the extensive flooding experienced by

[1] https://www.waterwheelfoundation.org

residents of New York and Vermont (wat, 2023). Donations are still being taken for this project as of the writing of this essay. In this research, we analyze the album titles, song titles and lyrics of Phish to consider their support for environmental protection and preservation. We believe that Phish rhetorically advocates for environmental protection in a range of ways that are both obvious and subtle. To test this theory, we use NLP techniques. We explored the application of topic modeling and machine learning technique, to dissect and interpret the environmental discourse embedded within Phish's lyrical compositions. Our research study tries to focus on the environmental topics of Phish, as well as to understand the popularity of these themes over time and location. To our knowledge, this study marks the first application of AI and NLP techniques for processing this lyrical information. Our objective is to offer comprehensive statistical insights into the presence of environmental themes in Phish's live performances, considering the dates and locations of their concerts. Our research is based on the utilization of NLP algorithms made feasible through data compiled from the Phish.net[2] lyrics collection created by the laborious work of Dr. Ellis Goddard, Associate Professor of Sociology at California State University, Northridge.

2 State of the Art

Several studies have explored Phish's relationship with their community and culture, covering topics such as public participation, copyright law, engagement, gender and racial diversity among fans, cultural practices of fans and music therapies (Carlson, 2020), (Kushner, 2020), (Marshall, 2003),(McClain, 2016), and (Rothstein, 2023). Additionally, ongoing research was gathered and discussed at the Phish Studies Conference, in 2019 and 2024 (Farzindar et al., 2024), organized by Oregon State University.

Our study focuses on Phish lyrics and song titles and includes both original Phish songs as well as cover songs written and performed by other bands. Phish regularly performs cover songs and the decision to include them is based on the fact that covers are a choice by the band. Consequently, they represent the musical and lyrical interests of the band as they perform unique songs by other artists.

Since lyrical texts consist of a sequence of words, multiple NLP methods could be applied to it. However, NLP methods are not always as effective for lyrics due to the differing nature of lyrics compared to traditional texts. These challenges have several reasons, such as creativity, ambiguity, variability, and emotional depth in texts. The text of lyrics often contains non-standard language, creative expressions, slang, and poetic devices. This makes it harder for NLP models, which are usually trained on more formal, standardized text, to interpret the meaning accurately. The presence of ambiguity, figurative language and metaphors are bold in lyrics, making it difficult for NLP models to accurately process the intended meaning. Lack of context and using very short sentences in lyrics are another leading factors in correct language processing and interpretation of emotional nuances.

Several studies explore themes in songs, often using LDA, a probabilistic topic modeling method (Liew et al., 2020). More recent work utilizes the Bidirectional Encoder Representations from Transformers (BERT), a pre-trained deep learning model by Google that generates word or sentence embeddings, capturing contextual and semantic meaning. Specific BERT models, such as MusicBERT, are tailored for NLP tasks involving both text and music (Rossetto and Dalton, 2020).

3 Methodology

In this research, we employ the topic modeling technique for lyrical information analysis. The goal is to identify clusters of words that frequently co-occur, aiming to represent topics related to the environment within the text and visualize them. For this purpose, we utilize topic modeling to classify Phish album titles, song titles, and lyrics as either "Environmental" or "Non-Environmental." Data classified as "Environmental" was subjected to further analysis, including mapping the time and location of live performances. The steps undertaken in this study consists of Web Scraping and Preprocessing, Topic Modeling, Environmental Classification, Evaluation and Visualizing. We utilized BERTopic, which captures nuanced semantic relationships between words and documents, leading to more accurate and interoperable topic clusters.

3.1 Web Scraping and Preprocessing

3.1.1 Web Scraping

Song lyrics and live performance data were scraped from Phish.net, providing a comprehensive dataset

[2]phish.net https://phish.net/

for analysis. Phish.net is an online community for fans of the Phish band, offering set lists, show reviews, and analysis of their music and performances.

Phish took a two-year hiatus starting in October 2000, resuming performances in December 2002, only to disband again in August 2004. This resulted in a gap in their concert data until their official reunion in March 2009, following an announcement in October 2008. Consequently, the dataset for Phish concerts is empty for the years 2005 through early 2009.

From 1980 to March 2024, a total of 1052 songs, with lyrics and performance data, was scraped from the website.

3.1.2 Exclusion of Instrumental Tracks

As noted earlier, our corpus includes both original Phish songs and songs Phish covers at their concerts. However, only songs with lyrics are included in the analysis, omitting instrumental tracks to focus solely on lyrical content. Instrumental tracks are songs that do not have any verbal lyrics and are completely consisting of instrumental music. After filtering out instrumental songs in our dataset, we were left with a total of 645 songs that contained full lyrics.

3.1.3 Tokenization and Data Processing

Data preprocessing included sentence tokenization and stop-word removal to ensure the quality and consistency of textual data. In the sentence tokenization process, each line of lyrics is segmented into individual sentences, breaking down the text into smaller units for analysis. This allows the BERTopic model, used in the next stage, to understand the context and meaning of each sentence independently, facilitating the clustering of similar topics or themes within the lyrics.

3.2 Topic Modeling

3.2.1 Embedding Generation

Embedding refers to the vector representations of words or sentences generated by pre-trained transformer models like BERT. These embeddings capture semantic meaning and context, allowing models such as BERTopic to analyze similarities between words or sentences based on their vector representations. By leveraging embeddings, BERTopic can cluster text data effectively, identifying topics or themes within the corpus.

In our study, the input for these models is the set of song lyrics, and output is the embeddings needed for the BERTopic models. The most popular models of Sentence Transformers for the English language were utilized to generate embeddings for the lyrics[3], namely model all- MiniLM-L6-v2 and model all-mpnet-base-v2.

3.2.2 Topic Modeling

BERTopic is a topic modeling tool that utilizes BERT embeddings to cluster documents based on their semantic similarity. Unlike traditional topic modeling methods like Latent Dirichlet Allocation (LDA), BERTopic captures nuanced semantic relationships between words and documents, resulting in more accurate and interpretable topic clusters.

The BERTopic model generated an output as a list of topics, with each list containing the key words highlighted within that topic. Each topic has its associated list of documents that represents that topic. In this study, we obtained a list of 34 topics generated from BERTopic model over the dataset of Phish lyrics.

3.3 Environmental Classification

An environmental classifier was employed to filter out topics related to environmental subjects from the total list of topics generated from BERTopic. The classifier used was: ESGBERT/EnvRoBERTa-environmental[4]. After classification, three topics were identified as containing environmental subjects from the list of 34 topics. We labeled the automatically selected topics from the topic modeling module, as described in the previous section, as **Water**, **Planets** and **Living things**. These topics include the following concepts:

- **Water**: rolls, away, water, sea, bouncing, flowing, wind, sky, room, light
- **Planets**: planet, slippin, flip, way, time, space, oh, world, soul, easy
- **Living things**: bug, hear, living, wind, quiet, sound, ringing, peeping, pane, frustration

Out of a total of 645 lyrics, 200 songs, making up approximately 31% of the corpus, were classified

[3] SentenceTransformers in Python framework for sentence, text and image embeddings https://huggingface.co/sentence-transformers?sort_models=downloads#models

[4] ESGBERT/EnvRoBERTa-environmental https://huggingface.co/ESGBERT/EnvRoBERTa-environmental

as belonging to environmental topics. This significant percentage highlights Phish's substantial engagement with environmental discourse throughout their music catalog.

3.4 Evaluation of models

3.4.1 Evaluation of topic modeling

For evaluating our topic modeling modules and select the best models, we used the C_V coherence score. The score is between $0 < x < 1$ and a higher score indicates that the top words in the topic frequently appear in similar contexts, suggesting that the topic is coherent and meaningful. Topic modeling for lyrics is challenging due to the poetic and metaphorical language used, which often conveys abstract themes rather than concrete topics, and words in lyrics can have multiple meanings or shift dramatically in context. However, in this study, a coherence score of C_V equal to 0.46 was obtained, indicating medium coherence. To further assess the medium coherence, we manually examined the three selected topics concerning environmental topics and concluded that this performance was sufficient to meet our objectives.

3.4.2 Evaluation of classifier

The automatic binary classifier analyzed 645 songs, classifying 194 as environmental. For evaluating the performance of classifiers, we needed a labeled dataset to check the precision of the machine's output, but we did not have any annotated data. For this purpose, we manually labeled a random sample of 143 lyrics as environmental and non-environmental. In addition to the classification task, we consider the confidence level of the classifier, which indicates its certainty about the predictions it makes.

In this study, considering an 80% confidence level in the classifier's labeling as environmental, the precision is 0.634 and recall is 0.866.

3.5 Visualization and Interpretation

To demonstrate the result of topic modeling techniques and automatic classification of topics, we develop an interactive visualization showing the distribution of topics related to the environment in Phish lyrics.

In our analysis of the Phish datasets, we employed several visualization techniques to gain insights. One approach involved mapping the time and location of the environmental songs' performed, with a specific focus on North America.

Additionally, we utilized statistics to track the occurrences of environmental songs over time and location. The Fig 1 revealed the percentage of environmental songs were shared with audiences across various cities hosted concert venues.

Furthermore, we calculated the trends of environmental themes in Phish live performances. Trends were calculated using a specific formula designed to measure changes relative to a baseline. Specifically, the trend for each value is calculated as follows:

$$\left(\frac{\text{Current value} - \text{Reference value}}{\text{Reference value}}\right) \times 100\%$$

This formula expresses the change as a percentage of the reference value, providing a standardized way to understand shifts over time or between datasets. In this analysis, the year 2019 was chosen as the reference value for trends over the years. The choice of 2019 as the baseline is strategic; it serves as a solid reference point because it is the most recent complete year before the disruptions caused by the COVID-19 pandemic. By using 2019 as the reference, it ensures that the data compared is from a period of relative normalcy, thereby providing a clear picture of how metrics have evolved from a pre-pandemic standpoint to the present. Fig 2 summarizes a comprehensive picture and confirms that the band has been increasingly active in promoting and addressing environmental topics through their concerts over years.

4 Conclusion

The analysis of Phish's lyrics using advanced natural language processing techniques reveals a strong and consistent presence of environmental themes in the band's music. By leveraging topic modeling, we identified clusters of lyrics focused on elements such as water, planets, and living things, demonstrating Phish's engagement with environmental discourse over time and across various locations. Our findings indicate that a significant percentage of Phish's lyrical content relates to environmental topics, highlighting the band's commitment to raising awareness through their music. This study provides a novel contribution to both musicology and environmental studies by using AI-driven techniques to quantify and visualize the influence of environmental themes in live performances. Moreover, the alignment between the band's concert locations and thematic content suggests that Phish

Figure 1: Percentage of Environmental songs vs location

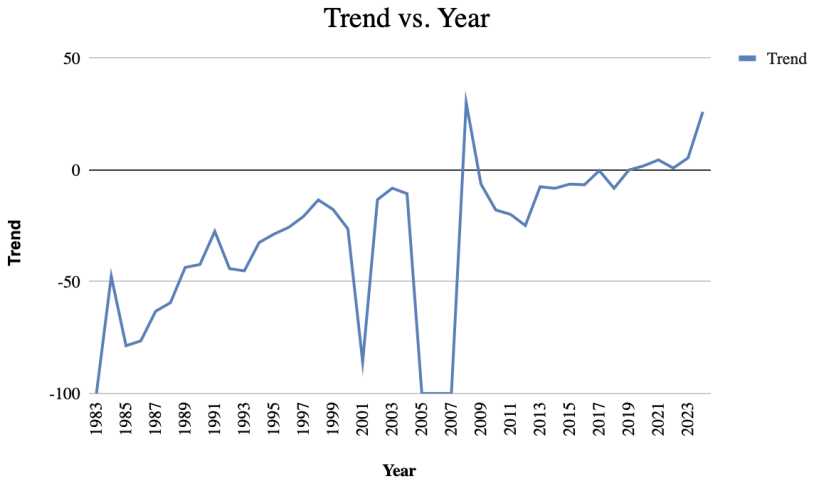


Figure 2: Trends of environmental themes in Phish live performances over Years

strategically incorporates environmental advocacy into their performances, further solidifying their role as a cultural force for environmental awareness. Future research could expand this framework to explore other thematic elements in their lyrics or apply similar methods to other artists and genres. The adaptability and scalability of these NLP techniques make them valuable tools for any study aimed at understanding the intersection of art, culture, and societal issues. As demonstrated in the Phish case study, the use of topic modeling, classification, and visualization can uncover latent themes in artistic works, providing insights into how artists communicate with their audiences and contribute to broader cultural movements.

References

2023. The WaterWheel Foundation - Phish concerts raise more than $3.5 million for flood relief in Vermont and New York.

Dennis Carlson. 2020. *A History of Progressive Music and Youth Culture: Phishing in America*, 1st edition edition. Peter Lang Inc., International Academic Publishers.

Anna Farzindar, Jason Jarvis, and Deepak Jayan. 2024. Divided sky: The environmental rhetoric of phish. In *Phish Studies Conference 2024*, Oregon State University, Corvallis, Oregon.

Eric Frankenberg. 2023. Top 10 Highest-Grossing Rock Tours of 2023.

Scott Kushner. 2020. Collecting and media change, or: Listening to Phish via app. *Convergence*, 26(4):969–989. Publisher: SAGE Publications Ltd.

Kongmeng Liew, Yukiko Uchida, Nao Maeura, and Eiji Aramaki. 2020. Classification of nostalgic music through LDA topic modeling and sentiment analysis of YouTube comments in Japanese songs. In *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, pages 78–82, Online. Association for Computational Linguistics.

Lee Marshall. 2003. For and Against the Record Industry: an Introduction to Bootleg Collectors and Tape Traders. pages 57–72.

Jordan M. McClain. 2016. Framing in Music Journalism: Making Sense of Phish's "Left-Field Success Story". *The Journal of Popular Culture*, 49(6):1206–1223. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jpcu.12489.

Eric Renner Brown. 2024. Phish Sphere Review: Jam Band Masters Four-Show Residency.

Federico Rossetto and Jeff Dalton. 2020. MusicBERT - learning multi-modal representations for music and text. In *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, pages 64–66, Online. Association for Computational Linguistics.

Caroline Rothstein. 2023. *I've Been Wading in the Whitest Sea: REFLECTIONS ON RACE, JUDAISM, AND PHISH*. Penn State Press. Google-Books-ID: bTrUEAAAQBAJ.

Michael Lee Simpson. 2024. Drew Carey Says Seeing Phish at Sphere Made U2 'Look Like a Bar Band'.

A Retrieval Augmented Approach for Text-to-Music Generation

Robie Gonzales[1] **Frank Rudzicz**[1,2]

[1]Dalhousie University [2]Vector Institute
{robie.gonzales,frank}@dal.ca

Abstract

Generative text-to-music models such as MusicGen are capable of generating high fidelity music conditioned on a text prompt. However, expressing the essential features of music with text is a challenging task. Furthermore, the limited set of text-music pairs leads to distributional shift, resulting in a consistent audio quality degradation with underspecified prompts. In this paper, we present a retrieval-augmented approach for text-to-music generation. We first pre-compute a dataset of text-music embeddings obtained from a contrastive language-audio pretrained encoder. Then, given an input text prompt, we retrieve the top k most similar musical aspects and augment the original prompt. This approach consistently generates music of higher audio quality as measured by the Frechét Audio Distance. We compare different retrieval strategies and find that augmented prompts dislay high text adherence Our findings show the potential for increased control in text-to-music generation.

1 Introduction

Modeling discrete representations of audio obtained from a neural audio codec has been an effective approach in tasks such as audio generation (Borsos et al., 2023; Kreuk et al., 2023), speech synthesis (Wang et al., 2023; Zhang et al., 2024) and self-supervised learning (Pepino et al., 2023). In particular, text-to-music generation (Copet et al., 2023; Agostinelli et al., 2023; Lam et al., 2023; Huang et al., 2023a; Schneider et al., 2024; Liu et al., 2023) has seen widespread adoption.

Despite their impressive capabilities, these models still suffer from distributional shifts, where underspecified user prompts lead to an audio quality degradation. Furthermore, constructing text prompts that accurately capture the user's creative intent while also expressing the essential features of music remains a challenge.

Inspired by the success of retrieval augmented generation (RAG) in natural language processing tasks, we present a retrieval augmented approach for text-to-music generation. While relatively simple, we show our approach consistently generates music of higher audio quality, while also displaying high text adherence. Our findings show potential for increased control in text-to-music generation.

2 Related Work

2.1 Music Generation

Recent generative music models can be roughly separated into two categories: transformer-based and diffusion-based models. MusicLM (Agostinelli et al., 2023) adopts a similar approach to AudioLM (Borsos et al., 2023), which represents audio using multiple streams of "semantic tokens" and "acoustic tokens" obtained from SoundStream (Zeghidour et al., 2021). MusicGen (Copet et al., 2023) adopts a single stage approach, where a transformer decoder is trained to predict multiple streams of discrete audio tokens using codebook interleaving patterns. MAGNeT (Ziv et al., 2024) extends this approach by introducing a masking schedule during training in which spans of tokens are predicted.

Conversely, diffusion models such as MeLoDy (Lam et al., 2023), Moûsai (Schneider et al., 2024), and AudioLDM (Liu et al., 2023) operate on learned, continuous representations of the audio signal. DITTO (Novack et al., 2024) and Music ControlNet (Wu et al., 2024) enable tailored music creation by directly optimizing control features in the latent space, whereas Mustango (Melechovsky et al., 2024) integrates textual metadata controls within the reverse diffusion step. In this work, we focus on the transformer based MusicGen (Copet et al., 2023).

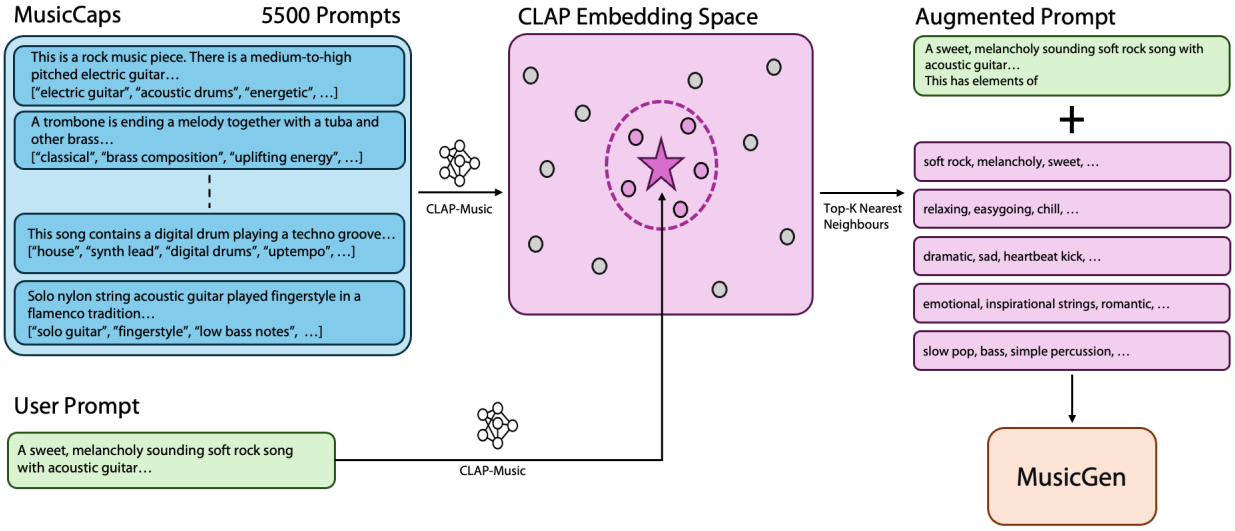


Figure 1: Overview of our retrieval augmented approach. We encode the text captions of MusicCaps using embeddings from CLAP. Given an input text prompt, we retrieve the top k most similar items. We extract their musical aspects list and concatenate them to the original prompt. This is fed as input to MusicGen for text-to-music generation.

2.2 Retrieval Augmented Generation

Retrieval augmented generation (RAG) has been a popular approach for integrating external knowledge from a retrieval module into a parametric language model, particularly for knowledge intensive tasks (Lewis et al., 2020). In this framework, contextually relevant documents from an external corpus are retrieved according to a query. This information is then augmented to the original input and guides the generation process. While there has been some work in applying RAG for general text-to-audio generation (Yuan et al., 2024) and speech (Wang et al., 2024), no work yet has focused on text-to-music generation.

3 Methodology

3.1 Dataset

For our experiments using retrieval and text-to-music generation, we use the MusicCaps dataset (Agostinelli et al., 2023) which consists of roughly 5500 text-music pairs. Each 10-second music clip is paired with a free-text caption describing the music (*This is a rock music piece*), and a list of musical aspects describing genre, mood, instrumentation, etc (*electric guitar, acoustic drums, energetic*).

3.2 Retrieval

The goal of the retrieval module is to retrieve a set of textual music aspects that are similar to the input text prompt. We first pre-compute a dataset of text embeddings obtained from a contrastive language-audio pre-trained encoder (CLAP) (Wu et al., 2023). We use the `music-audio` checkpoint [1], which is trained on AudioSet (Gemmeke et al., 2017), LAION-Audio-630k (Wu et al., 2023), and a dataset of music samples. We encode each free-text caption in MusicCaps which results in a fixed sized 512-dimensional embedding. To store these embeddings, we use Spotify's Annoy [2], an approximate nearest neighbour search library.

Given an input text prompt, we retrieve the top k most similar captions ranked by Euclidean distance. We then extract their musical aspects and perform preprocessing to remove duplicates and words that signal low quality (*low quality, poor audio quality, amateur recording*). Finally, we combine all retrieved musical aspects, prefix them with *"This has elements of "* and concatenate them to the original prompt.

We experiment with various retrieval strategies. We vary the number of retrieved items using $k = 3, 5, 10$. We also experiment with retrieving similar items using the CLAP text embedding of the musical aspect list, as well as retrieving random musical aspects.

[1] https://huggingface.co/lukewys/laion_clap/blob/main/music_audioset_epoch_15_esc_90.14.pt

[2] https://github.com/spotify/annoy

Method	$\text{FAD}_{\text{CLAP-Audio}} \downarrow$	$\text{FAD}_{\text{CLAP-Music}} \downarrow$	$\text{FAD}_{\text{VGGish}} \downarrow$	KL $\downarrow$	$\text{CLAP}_{\text{score}} \uparrow$
Unconditional	0.4668 $\pm$ 0.0061	0.5063 $\pm$ 0.0033	7.1027 $\pm$ 0.0048	2.1013 $\pm$ 0.0027	-
First Aspect	0.4055 $\pm$ 0.0024	0.4401 $\pm$ 0.0009	5.3287 $\pm$ 0.0029	2.0567 $\pm$ 0.0113	0.1038 $\pm$ 0.0063
First Sentence	0.3520 $\pm$ 0.0031	0.4055 $\pm$ 0.0013	4.8330 $\pm$ 0.0024	1.5212 $\pm$ 0.0031	0.0735 $\pm$ 0.0054
Full Caption	0.3443 $\pm$ 0.0026	0.4027 $\pm$ 0.0007	4.7895 $\pm$ 0.2830	1.3044 $\pm$ 0.0011	0.0997 $\pm$ 0.0113
Caption's Nearest Neighbours					
Augmented ($k = 3$)	0.3363 $\pm$ 0.0021	0.4093 $\pm$ 0.0009	4.8390 $\pm$ 0.0021	1.3485 $\pm$ 0.0017	**0.2863** $\pm$ 0.0026
Augmented ($k = 5$)	**0.3189** $\pm$ 0.0014	**0.3878** $\pm$ 0.0004	**4.3458** $\pm$ 0.0028	1.2556 $\pm$ 0.0014	0.2810 $\pm$ 0.0121
Augmented ($k = 10$)	0.3496 $\pm$ 0.0130	0.4116 $\pm$ 0.0012	4.6627 $\pm$ 0.0017	1.3538 $\pm$ 0.0013	0.2854 $\pm$ 0.0070
Aspect's Nearest Neighbours					
Augmented ($k = 3$)	0.3519 $\pm$ 0.0049	0.4187 $\pm$, 0.0102	5.1123 $\pm$ 0.0007	1.2776 $\pm$ 0.0059	0.2756 $\pm$ 0.0024
Augmented ($k = 5$)	0.3407 $\pm$ 0.0068	0.4131 $\pm$ 0.0087	4.9794 $\pm$ 0.0076	1.3611 $\pm$ 0.0112	0.2790 $\pm$ 0.0015
Augmented ($k = 10$)	0.3507 $\pm$ 0.0178	0.4095 $\pm$ 0.0100	4.8538 $\pm$ 0.0390	**1.2272** $\pm$ 0.0011	0.2688 $\pm$0.0032
Random Aspects					
Augmented $k = 3$)	0.3640 $\pm$ 0.0033	0.4403 $\pm$ 0.0018	5.8419 $\pm$ 0.0006	1.6052 $\pm$ 0.0042	0.2453 $\pm$ 0.0031
Augmented ($k = 5$)	0.4433 $\pm$ 0.0270	0.4511 $\pm$ 0.0024	6.1379 $\pm$ 0.0041	1.7267 $\pm$ 0.1520	0.2358 $\pm$ 0.0026
Augmented ($k = 10$)	0.4159 $\pm$ 0.0198	0.4801 $\pm$ 0.0028	6.6805 $\pm$ 0.0023	1.5736 $\pm$ 0.0371	0.2199 $\pm$ 0.0033

Table 1: Quantitative evaluation results. Mean values and 95% confidence intervals are reported. For the augmented caption, we experiment with retrieving the caption's nearest neighbours, the musical aspect's nearest neighbours and random aspects. A low FAD score indicates the generated music is plausible. A low KL score indicates the generated music shares similar concepts with the reference set. A high CLAP score indicates the generated music adheres to the text prompt.

3.3 Text-to-Music Generation

An audio language model is composed of two components: (i) a compression model, which handles a mapping between audio signals and discrete audio tokens, and (ii) a transformer decoder language model, which operates on these audio tokens. To facilitate text conditioning, a pre-trained text encoder is integrated into the cross-attention blocks of the transformer decoder.

Given a discrete representation of the audio signal z, our goal is to model the joint probability distribution $p_\theta(z \mid y)$, where y is a semantic representation of the condition. This can be computed as a product of its conditional probabilities:

$$p_\theta\left(z_1, \ldots, z_n \mid y\right) = \prod_{i=1}^{n} p_\theta\left(z_i \mid z_1, \ldots, z_{i-1}, y\right) \tag{1}$$

In this work, we are interested in the effect of augmenting y with relevant musical information and how it affects the generation process.

We generate baseline music samples using several methods: (1) no text prompt (unconditional), (2) using only the first musical aspect in the list, (3) using only the first sentence in the text caption, (4) the full text caption.

3.4 Evaluation

Evaluating generative music models remains a challenge (Gui et al., 2024). Given we are interested in the effect of an augmented prompt in text-to-music generation, we aim to capture two important aspects: the audio quality and the adherence to the text description.

Frechét Audio Distance (FAD) The Frechét Audio Distance (Kilgour et al., 2019) is a reference-free audio quality metric which correlates well with human perception. The FAD is computed by comparing a reference set of audio samples to an evaluation set in terms of their distributions in an embedding space. A low FAD score indicates the audio of the evaluation set is plausible. We use the FAD toolkit [3] to evaluate our generated music samples in three embedding spaces: $\text{CLAP}_{\text{Audio}}$, $\text{CLAP}_{\text{Music}}$ and VGGish.

Kullback-Leiber Divergence (KL) The KL-divergence is a measure of how one probability distribution diverges from a second, expected probability distribution. Due to the complex nature of music, there is a many-to-many relationship between text descriptions and music clips. Therefore, we use a classifier (Koutini et al., 2022) trained for

[3] https://github.com/microsoft/fadtk

multi-label classification on AudioSet to compute the KL-divergence over the class probabilities between the reference set and generated music. The generated music is expected to share similar concepts with the reference set when the KL is low.

CLAP Score As a joint text-audio embedding model, CLAP can be used to quantify the similarity between text-audio pairs. We compute both the text embedding $f_{text}(\cdot)$ of the text caption, and the audio embedding $f_{audio}(\cdot)$ of the generated music sample. Similar to the MuLan Cycle Consistency (MCC) (Huang et al., 2022), the CLAP score is then defined as the average cosine similiarity between these embeddings (Copet et al., 2023; Huang et al., 2023b). A high CLAP score indicates the generated music adheres to the text prompt.

4 Results

Table 1 presents the results of the RAG approach against baseline generation methods. In general, FAD scores computed in the $\text{CLAP}_{\text{Audio}}$ and $\text{CLAP}_{\text{Music}}$ embedding spaces are lower than the VGGish embedding space. This could be attributed to VGGish being trained on a classification task at 16kHz, while CLAP is trained on a contrastive task at 48kHz. The higher dimensional CLAP features may also capture more complex musical features.

For the baseline generation methods, unconditional generation consistently generates music of poor audio quality. Specifying a single musical aspect or first sentence of the caption improves quality, while the full text caption achieves the best scores. This suggests that the conditioning text encoder plays a key role in influencing the generation process. The CLAP score is highest when specifying a single musical aspect, likely because matching the audio to a single word presents a simpler task than aligning with a more complex sentence.

For the RAG methods, retrieving similar items based on the caption outperforms retrieving similar items based on the musical aspects. This is reasonable as the musical aspects aim to capture more qualitative features and as a result could diverge more from the intended caption description. Retrieving five similar items based on the caption achieves the best FAD scores, suggesting a trade-off with prompt length. Despite their relevance, too many aspect qualities may hinder the models performance instead of focusing on a select few. The best CLAP score is achieved by retrieving three similar items based on the caption, which aligns with the notion that matching fewer relevant words presents a simpler task. Interestingly, the best KL score is achieved by retrieving ten similar items based on the musical aspect list. This could be due to our implementation of calculating KL, where we use a classifier trained for a multi-label task on AudioSet. By retrieving many diverse aspects, we increase the probability of matching with multiple labels.

Finally, retrieving random items generates music of worse or comparable audio quality to baseline generation methods. Again, this suggests a trade-off with the prompt length and relevance. This also demonstrates the ability of the pre-trained text encoder to transfer useful representations when generating diverse music.

5 Future Work

In this work, we explored the overall effect of an augmented prompt when generating music with MusicGen. However, it would be more valuable to investigate specifically how the augmented tokens affect the internal representations. SMITIN (Koo et al., 2024) trains classifier probes to identify self-attention heads that perform instrument recognition. Then, they introduce an inference time intervention technique for steering the generated output towards the desired musical trait. Extending this approach to other specific control methods for various musical features is a desirable goal.

People value agency and control over creative collaborations with generative AI models. As such, we want to build systems that promote interactive, human-centered approaches. Equipping MusicGen with the ability to refine and build upon previously generated output is another valuable direction.

6 Conclusion

In this paper, we presented a retrieval augmented approach for text-to-music generation. While relatively simple, we show our method consistently generates music of higher audio quality while displaying high text adherence. We compare the trade-offs of various retrieval strategies and suggest extensions to this work. Our findings show the potential for increased control in text-to-music generation.

7 Ethics Statement

Large scale generative models raises questions regarding ethics and societal consequences of their use. Generative text-to-music models can represent an unfair competition for artists which is an open problem. Another potential bias is the lack of diversity in the MusicCaps dataset, which contains a larger proportion of Western music. Through open research, we hope that such generative models can become useful as a tool for amateur musicians and professionals.

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.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023. Audiolm: A language modeling approach to audio generation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:2523–2533.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Defossez. 2023. Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, volume 36, pages 47704–47720. Curran Associates, Inc.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. 2024. Adapting frechet audio distance for generative music evaluation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1331–1335.

Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P. W. Ellis. 2022. Mulan: A joint embedding of music audio and natural language. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022*, pages 559–566.

Qingqing Huang, Daniel S. Park, Tao Wang, Timo I. Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, Jesse Engel, Quoc V. Le, William Chan, Zhifeng Chen, and Wei Han. 2023a. Noise2music: Text-conditioned music generation with diffusion models. *Preprint*, arXiv:2302.03917.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023b. Make-an-audio: text-to-audio generation with prompt-enhanced diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2019. Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms. In *Proc. Interspeech 2019*, pages 2350–2354.

Junghyun Koo, Gordon Wichern, Francois G. Germain, Sameer Khurana, and Jonathan Le Roux. 2024. Smitin: Self-monitored inference-time intervention for generative music transformers. *Preprint*, arXiv:2404.02252.

Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. 2022. Efficient training of audio transformers with patchout. In *Interspeech 2022*, pages 2753–2757.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2023. Audiogen: Textually guided audio generation. In *The Eleventh International Conference on Learning Representations*.

Max W. Y. Lam, Qiao Tian, Tang Li, Zongyu Yin, Siyuan Feng, Ming Tu, Yuliang Ji, Rui Xia, Mingbo Ma, Xuchen Song, Jitong Chen, Wang Yuping, and Yuxuan Wang. 2023. Efficient neural music generation. In *Advances in Neural Information Processing Systems*, volume 36, pages 17450–17463. Curran Associates, Inc.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. AudioLDM: Text-to-audio generation with latent diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 21450–21474. PMLR.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2024. Mustango: Toward controllable text-to-music generation. In *Proceedings of the 2024*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8293–8316, Mexico City, Mexico. Association for Computational Linguistics.

Zachary Novack, Julian Mcauley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan. 2024. DITTO: Diffusion inference-time t-optimization for music generation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 38426–38447. PMLR.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2023. Encodecmae: Leveraging neural codecs for universal audio representation learning. *Preprint*, arXiv:2309.07391.

Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. 2024. Moûsai: Efficient text-to-music diffusion models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8050–8068, Bangkok, Thailand. Association for Computational Linguistics.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *CoRR*, abs/2301.02111.

Mingqiu Wang, Izhak Shafran, Hagen Soltau, Wei Han, Yuan Cao, Dian Yu, and Laurent El Shafey. 2024. Retrieval augmented end-to-end spoken dialog models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12056–12060.

Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J. Bryan. 2024. Music controlnet: Multiple time-varying controls for music generation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 32:2692–2703.

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 *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yi Yuan, Haohe Liu, Xubo Liu, Qiushi Huang, Mark D. Plumbley, and Wenwu Wang. 2024. Retrieval-augmented text-to-audio generation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 581–585.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 30:495–507.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024. Speechtokenizer: Unified speech tokenizer for speech language models. In *The Twelfth International Conference on Learning Representations*.

Alon Ziv, Itai Gat, Gael Le Lan, Tal Remez, Felix Kreuk, Jade Copet, Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi. 2024. Masked audio generative modeling. In *The Twelfth International Conference on Learning Representations*.

Information Extraction of Music Entities in Conversational Music Queries

Simon Hachmeier and Robert Jäschke
School of Library and Information Science
Humboldt-Universität zu Berlin
{simon.hachmeier,robert.jaeschke}@hu-berlin.de

Abstract

The detection of music entities such as songs or performing artists in natural language queries is an important task when designing conversational music recommendation agents. Previous research has observed the applicability of named entity recognition approaches for this task based on pre-trained encoders like BERT. In recent years, large language models (LLMs) have surpassed these encoders in a variety of downstream tasks. In this paper, we validate the use of LLMs for information extraction of music entities in conversational queries by few-shot prompting. We test different numbers of examples and compare two sampling methods to obtain few-shot examples. Our results indicate that LLM performance can achieve state-of-the-art performance in the task.

1 Introduction

Detecting music entities such as songs or musical artists in natural language queries is a key component of conversational music agents (Jannach et al., 2021). In such queries, users request music entities they want to listen to in a conversational way as an alternative to traditional text search.

The task of detecting music entities is typically modeled as named entity recognition (NER) which was earlier addressed by probabilistic approaches (Liljeqvist, 2016; Porcaro and Saggion, 2019). More recently, pre-trained encoders demonstrated strong performance in the NER task in the music domain (Xu and Qi, 2022; Epure and Hennequin, 2023).

With the advent of large language models (LLMs) such as GPT-3.5 for text generation tasks, these are increasingly used for NER and the related task of information extraction (IE) (Wang et al., 2023a; Ashok and Lipton, 2023; Zhang et al., 2023). Several studies found that encoder-only models still outperform LLMs (Wang et al., 2023b; Ma et al., 2023; Sun et al., 2023; Zhou et al., 2024). However, LLMs are usually trained on much larger datasets than encoder-only models which theoretically makes these more likely to capture music knowledge to some extent.

In this paper, we investigate the success of LLMs for IE of music entities in conversational music queries (e.g., *give me some artists like metallica*).

We prompt LLMs to label utterances in the queries with respective labels (*title* and *artist*) and compare these to two strong baseline encoder-only models. We investigate the difference of two few-shot sampling methods and different numbers of few-shot examples. Lastly, we outline some contextual cues captured by the best performing LLM which we request in our prompt to reflect internal reasoning. We release our code publicly.[1]

In the next section, we outline related work in IE and NER using LLMs. In Section 3 we describe our proposed method. In Section 4 we describe our used dataset and baselines before presenting the results in Section 5. Lastly, we close this paper with Section 6.

2 Related Work

In this section we outline related work in IE and NER with LLMs. NER is a subtask of IE in which a sequence of labels per token or character is obtained. The more general task of IE refers to the extraction of relevant information in some structured form, but not necessarily a sequence of labels and possibly with additional steps (e.g., normalization). Research towards the use of LLMs for IE comprises the direct use of LLMs for the task (e.g., by instruction tuning) or auxiliary use in combination with encoder-only models (e.g., BERT).

A line of research has validated the usefulness of LLMs for NER. Beside Li et al. (2023) which fine-tune Llama-2, the most prevalent strategy appears to be prompting the LLM (Wang et al., 2023a;

[1] https://github.com/progsi/YTUnCoverLLM

Ashok and Lipton, 2023; Jung et al., 2024). Jung et al. (2024) relies on a single prompt without few-shot examples. Wang et al. (2023a) prompt GPT-3 and provide few-shot examples retrieved by a nearest neighbor search. Their approach achieves a performance close to the state-of-the-art based on BERT (Devlin et al., 2018). Ashok and Lipton (2023) demonstrate that GPT-3.5 and GPT-4 beat other LLMs such as T5XXL in the task. The authors include entity label descriptions and request reasoning for predicted entities. Sun et al. (2023) state that LLMs underperform in comparison to encoder-only models due to reasons like hallucination and context limits in few-shot settings. They provide various means to overcome this issue, such as demonstration retrieval and self-verification.

The mixed performance of LLMs for IE and NER motivates another paradigms which favors the auxiliary use of LLMs together with encoder-only models trained in a supervised fashion.

Zhang et al. (2024) argue that the lack of speciality of LLMs is a major factor. Hence, they propose an approach to combine those with encoder-only models, utilizing the LLM only for relabeling of initially uncertain predictions. Similarly, Ma et al. (2023) state that LLMs are better to use for hard samples than for general use in IE. They suggest to rather use LLM to re-rank of outputs obtained from a pre-trained encoder-only model. Other works include the auxiliary LLM purely for data augmentation Ye et al. (2024) or model distillation (Zhou et al., 2024; Peng et al., 2024). In the following, we propose our approach for IE from music queries using LLMs.

3 Music Entity Extraction with LLMs

The goal of our approach is the extraction of music entities from natural language queries. When users request music recommendations, the query can refer to various aspects such as genres, moods, titles (e.g., album or song titles) or performing artists (performers). We only focus on musical entities such as songs and albums (represented by their title) and performing artists (represented by their name). For each query, we aim to extract all the entities of this type and permit the possibility of no relevant entities being contained. Thus, the query *recommend me some rock songs* should yield no result, while the query *something similar to metallica st anger* should yield the utterance *metallica* with the label *performer* and the utterance *st anger* with the label *title* which refers to the American band *Metallica* and their album *St. Anger*. We model this task as an IE problem as we explain in the following.

Instruction
From the following text, which contains a user request for music suggestions, extract all the relevant entities that you find.

Entity Attributes
- **utterance**: The utterance of the entity in the text. For example "the beatles" in "recommend me music like the beatles". An utterance can only be of a type for which labels are defined.
- **label**: The label of the entity. It can either be 'TITLE' (if the utterance refers to a song or album name), 'PERFORMER' (if the utterance refers to a performing artist) or 'OTHER' for any other entity type.
- **cue**: The contextual cue which indicates the entity (e.g., "music like" in "recommend me music like the beatles" indicating "the beatles")

Examples
Input: stuff like flylo
({'utterance': 'flylo', 'label': 'performer', 'cue': ''})
Input: dré anthony brand new
...
Output Schema

```
from pydantic import BaseModel

class MusicEntity(BaseModel):
    """
    Data model of a music entity
    """
    utterance: str
    label: str
    cue: str
```

Input
songs similar to black bird by alter bridge

Figure 1: Prompt with few-shot examples and input text.

Instruction To obtain a structured output from an LLM, we define a Pydantic (Colvin et al., 2023) output schema and detailed instruction (cf. Figure 1). Previous approaches for IE with LLMs have discovered the importance of detailed attribute explanations with examples (Wang et al., 2023a; Ashok and Lipton, 2023; Zhang et al., 2023). Thus, we include explanations for each of the attributes. We also inclue a wildcard label *other* which we found helpful to improve the precision of LLMs (see Section 4). Beside utterance and label attributes, we request contextual cues.

Contextual Cues In theory, one can identify music entities in text by two ways. First, one can

simply rely on world knowledge. This way, even in difficult cases, one can identify a music entity with a correct label. For instance, the query containing *metallica st anger* which we showed earlier does not contain any cue clarifying that these are two entities and more precisely that *metallica* refers to a performer and *st anger* to an album. In contrast, some queries indicate the entity labels more clearly. For example *songs like nothing else matters by metallica* contains the cues *songs* and *by* which indicate the relationship *[song] by [performer]*. To gather more insights behind the internal reasoning of LLMs, we include an attribute *cue* in the structured output which should capture the contexts from the queries. This idea resembles the explanations requested in the prompt by Ashok and Lipton (2023).

Few-Shot Additionally to the zero-shot approach we described, we experiment with few-shot settings. We construct a few-shot example dataset which is the same dataset as the training dataset of the baseline models (cf. Section 4). We experiment with different numbers of k, corresponding to the amount of sampled examples at each iteration. Since the annotated dataset does not include contextual cues, we omit those in the few-shot examples. Beside random sampling, we experiment with a sampling approach that relies on the most similar k items from the example dataset, similar to the nearest neighbor approach by Wang et al. (2023a), but we use term frequency inverse document frequency (tf-idf) vectors and the Cosine similarity as a metric. To not let the actual title and performer strings impact the similarity, we replace them in the examples by respective masks. For instance, in a example query *songs like nothing else matters by metallica* we obtain *songs like [song] by [performer]*.

4 Experimental Design

4.1 Implementation Details

We test different parameter values of $k \in \{0, 5, 15, 25, 35, 45\}$. We tested different LLMs for their capability to output structured content reliably. For example, we tested Llama-3-8B[2] but it failed too often to conform to the output structure. For our experiments, we use the following three LLMs:

GPT-3.5-Turbo: An LLM that supports function calling and is well suited for structured output.[3] We use *gpt-3.5-turbo-0125*.[4]

Mistral-7B: An open-source LLM by Jiang et al. (2023) suitable for structured output without function calling.

Mixtral-8x7B: An open-source LLM following the mixture of experts (MoE) paradigma (Jiang et al., 2024).

We also experimented with the use of the label *other* and compared the precision for Artist and WoA respectively. While Artist precision was relatively stable, we observed a decrease of 0.27 for WoA precision. That is, because a lot of more generic utterances like genres or moods were detected as WoAs. Thus, we decided to use the *other* label for all for all further experimental runs and we simply ignore the respective outputs to compute the WoA and Artist evaluation metrics.

4.2 Dataset & Baselines

We use the MusicRecoNER (Epure and Hennequin, 2023) dataset which is based on a subreddit[5] in which users request music suggestions by mentioning reference entities of the type performing artists or other entities such as song titles or music albums (labeled *title*).[6] The dataset is split into four subsets, three with 600 and one with 751 queries. On average, each query has two entity mentions but around 56% queries do not have any entity mention. We fine-tune two strong baselines and report the results using 4-fold cross validation as done by Epure and Hennequin (2023):

BERT (Devlin et al., 2018): A bi-directional encoder pre-trained by cloze tasks such as masked language modeling. It achieves comparable performance to MPNet (Song et al., 2020) for the task.

RoBERTa (Liu et al., 2019): This encoder has the same architecture as BERT but was pre-trained using a different training scheme. The model surpasses vanilla BERT on a variety of downstream tasks.

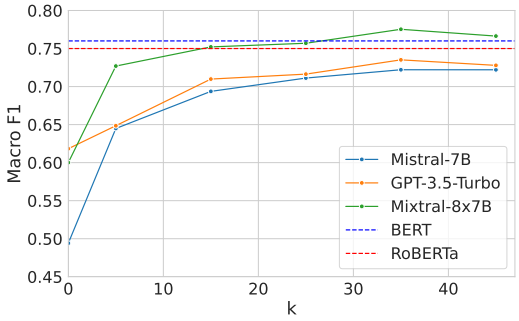
[2] see https://llama.meta.com/llama3/

[3] see OpenAI Function Calling Guide

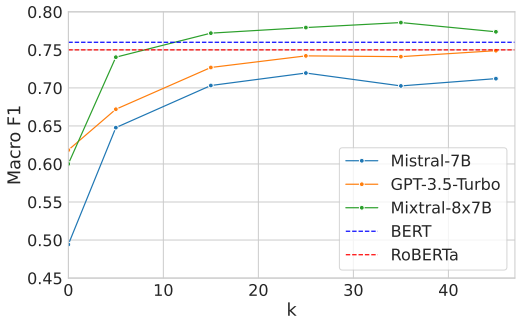
[4] see OpenAI Models

[5] www.reddit.com/r/musicsuggestions/

[6] Please note that we renamed the label in our prompts, since we found that the LLMs performance increased when using *title* instead of the original name *Work of Art* or *WoA*.



(a) F1 Scores for random few-shot sampling.



(b) F1 Scores tf-idf few-shot sampling.

Figure 2: Comparison of F1 Scores under strict evaluation scheme for different methods.

5 Results

In this section, we present the results of our previously presented experiments. All of the results are obtained with 4-fold cross validation using the split from Epure and Hennequin (2023).

In Figure 2 we show F1 scores as a function of k for both sampling methods of few-shot examples. The effect of tf-idf sampling as opposed to random sampling seems to have no positive effect on the performance of Mistral-7B and just a minor effect on GPT-3.5-Turbo. Mixtral-8x7B is the best performing LLM and achieves higher F1 scores than the baselines for $k = 35$ and random sampling. Using tf-idf sampling, it exceeds the baseline for smaller values of $k = 15$ and $k = 25$ and it achieves the highest F1 score in the experiment at close to 0.80. However, the performance for $k = 45$ decreases which is also the case for the other models at tf-idf sampling. At random sampling, the performance appears to stagnate for $k \geq 35$ as well, but experiments with even higher values are necessary to fully exploit the potential of even more examples.

To gather more detailed insights in LLM performance against the baselines, we report the precision and recall of both entity labels in Table 1. While BERT has the highest recall for both labels, it has a lower precision by a substantial margin compared to GPT-3.5-Turbo and Mixtral-8x7B for performers. Apparently, the recognition of titles in the queries is a more difficult task, since all the models undershoot both metrics compared to performers.

	Perf.		**Title**	
	Pr	Re	Pr	Re
BERT	0.81	**0.82**	0.72	**0.77**
RoBERTa	0.78	0.78	0.72	0.75
GPT-3.5-Turbo	**0.91**	0.78	0.64	0.65
Mistral-7B	0.78	0.71	0.74	0.61
Mixtral-8x7B	0.89	0.78	**0.77**	0.72

Table 1: Precision (Pr) and recall (Re) per labels performer and title for LLMs with $k = 35$ examples against the baselines.

Lastly, we investigate the contextual cues returned by the two best models: Mixtral-8x7B and GPT-3.5-Turbo with $k = 35$. We observe that cues indicating WoAs are less effective, resulting in 0.54 and 0.43 of WoA precision respectively. In contrast, the respective Artist precision is higher with 0.71 and 0.78. Frequent successful cues of Mixtral-8x7B are *ft*, *featuring* and *remix*. It is noteworthy, that Mixtral-8x7B and GPT-3.5-Turbo only returned cues in around 15% of cases, which might be due to the absence of cues in the few-shot examples.

6 Conclusion and Limitations

In this paper, we performed IE of music entities in conversational music queries with three LLMs as an alternative to the previously suggested NER with encoder-only models like BERT. We showed that tf-idf sampling to obtain similar few-shot examples to the query text can enhance the LLM performance, especially in case of the best performing model Mixtral-8x7B. The observed increase in F1 is mostly achieved by improved precision which leads to an overall improvement against the baselines. In future work, the inclusion of annotations for contextual cues could be helpful to encourage the LLMs to return those more frequently and possibly encourage better reasoning. Further, our study motivates experiments with even more capable LLMs such as Llama-3-70B.

References

Dhananjay Ashok and Zachary C Lipton. 2023. Promptner: Prompting for named entity recognition. *arXiv preprint arXiv:2305.15444*.

Samuel Colvin, Eric Jolibois, Hasan Ramezani, Adrian Garcia Badaracco, Terrence Dorsey, David Montague, Serge Matveenko, Marcelo Trylesinski, Sydney Runkle, David Hewitt, and Alex Hall. 2023. Pydantic. If you use this software, please cite it as below.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Sung Jae Jung, Hajung Kim, and Kyoung Sang Jang. 2024. Llm based biological named entity recognition from scientific literature. In *2024 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 433–435. IEEE.

Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu-lee Wang, Qing Li, and Xiaoqin Zhong. 2023. Label supervised llama finetuning. *arXiv preprint arXiv:2310.01208*.

Sandra Liljeqvist. 2016. Named entity recognition for search queries in the music domain.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

Letian Peng, Zilong Wang, Feng Yao, Zihan Wang, and Jingbo Shang. 2024. Metaie: Distilling a meta model from llm for all kinds of information extraction tasks. *arXiv preprint arXiv:2404.00457*.

Lorenzo Porcaro and Horacio Saggion. 2019. Recognizing musical entities in user-generated content. *Computación y Sistemas*, 23(3):1079–1088.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Preprint*, arXiv:2004.09297.

Xiaofei Sun, Linfeng Dong, Xiaoya Li, Zhen Wan, Shuhe Wang, Tianwei Zhang, Jiwei Li, Fei Cheng, Lingjuan Lyu, Fei Wu, and Guoyin Wang. 2023. Pushing the limits of chatgpt on nlp tasks. *Preprint*, arXiv:2306.09719.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023a. Gpt-ner: Named entity recognition via large language models. *Preprint*, arXiv:2304.10428.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023b. Instructuie: multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Wenjia Xu and Yangyang Qi. 2022. Gazetteer enhanced named entity recognition for musical user-generated content. In *2022 3rd International Conference on Computer Science and Management Technology (ICCSMT)*, pages 40–43. IEEE.

Junjie Ye, Nuo Xu, Yikun Wang, Jie Zhou, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llm-da: Data augmentation via large language models for few-shot named entity recognition. *arXiv preprint arXiv:2402.14568*.

Mozhi Zhang, Hang Yan, Yaqian Zhou, and Xipeng Qiu. 2023. Promptner: A prompting method for few-shot named entity recognition via k nearest neighbor search. *arXiv preprint arXiv:2305.12217*.

Zhen Zhang, Yuhua Zhao, Hang Gao, and Mengting Hu. 2024. Linkner: Linking local named entity recognition models to large language models using uncertainty. In *Proceedings of the ACM on Web Conference 2024*, WWW '24, page 4047–4058, New York, NY, USA. Association for Computing Machinery.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. Universalner: Targeted distillation from large language models for open named entity recognition. *Preprint*, arXiv:2308.03279.

Leveraging User-Generated Metadata of Online Videos for Cover Song Identification

Simon Hachmeier and Robert Jäschke
School of Library and Information Science
Humboldt-Universität zu Berlin
{simon.hachmeier,robert.jaeschke}@hu-berlin.de

Abstract

YouTube is a rich source of cover songs. Since the platform itself is organized in terms of videos rather than songs, the retrieval of covers is not trivial. The field of cover song identification addresses this problem and provides approaches that usually rely on audio content. However, including the user-generated video metadata available on YouTube promises improved identification results. In this paper, we propose a multi-modal approach for cover song identification on online video platforms. We combine the entity resolution models with audio-based approaches using a ranking model. Our findings implicate that leveraging user-generated metadata can stabilize cover song identification performance on YouTube.

1 Introduction

Music is a popular content category on YouTube (Montero and Mora-Fernandez, 2020). Uploaders share music in a variety of contexts, ranging from amateur covers to mashups (Airoldi et al., 2016; Liikkanen and Salovaara, 2015). Since YouTube is not organized in terms of songs but rather in terms of online videos,[1] finding cover versions of songs is a non-trivial retrieval task.

Driven by applications such as copyright infringement detection, cover song identification (CSI) deals with the retrieval of covers. The key challenge of CSI is to compare songs based on properties which can indicate their association (e.g., melody, lyrics) while discarding irrelevant information (e.g., timbre). Consequently, current research efforts are mainly audio-based (Du et al., 2023; Liu et al., 2023; Hu et al., 2022; Yu et al., 2020) with limited consideration of non-audio features such as lyrics (Abrassart and Doras, 2022). However, the utilization of user-generated metadata of online videos like in similar tasks (Agrawal and Sureka; Smith et al., 2017), has not yet been considered in CSI.

We propose to model CSI on online video platforms (OCSI) as a multi-modal problem, based on the hypothesis that uploaders tend to describe their videos using attributes of songs (e.g., song title, performer name) to make them easily findable. In this work, we propose multi-modal ensembles combining entity resolution models with audio-based CSI models in a late-fusion fashion using the ranking model LambdaMART (Wu et al., 2010). We compare the performance of the proposed ensembles with the performance of the CSI models. Further, we study the robustness of ER approaches in difficult cases such as song title variations or hard negatives on YouTube and provide our code and results.[2]

2 Multi-Modal Online Cover Song Identification

The items in our task are online videos. The goal of OCSI is the retrieval of items associated with the same musical work as a query item. In traditional CSI, only representations concerning musical content (e.g., audio and lyrics) are considered. We additionally leverage user-generated metadata. In Figure 1 we show an example of a query-candidate pair.

Each item is represented by attributes derived from its audio data and attributes from its user-generated metadata (video title, channel name, video description, and a set of keywords). For simplicity, we assume that each item contains only one song. A song has the attributes song title, performer name, and a work identifier. Songs which are associated with the same musical work are considered relevant in the retrieval scenario when one song is used as a query to retrieve the other song.

We model the task of OCSI as a multi-modal

[1] Except for YouTube's streaming service *YouTube Music*.

[2] https://github.com/progsi/er_csi

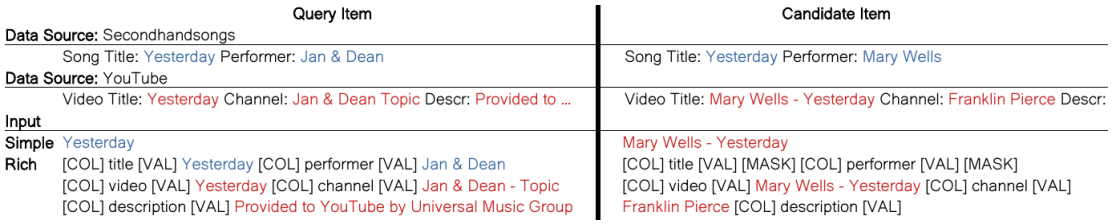


Figure 1: Example of the input of items of the work "Yesterday" written by John Lennon and Paul McCartney. Colors in the box frames and text indicate the data source: blue stands for Secondhandsongs and red for YouTube.

problem involving metadata and audio representations. Our overall system will consist of modules to compute the pairwise similarities for each of both modalities. Then, a ranking model combines both outputs to compute an overall rank. In the following, we explain our entity-resolution (ER) methods used to model similarities in the metadata domain.

2.1 Entity Resolution

Fuzzy Matching We use the token ratio function from rapidfuzz (Bachmann, 2021), which turned out to be the best performing fuzzy matcher for the task in a preliminary experiment. For a pair of strings, the function returns the maximum between their normalized Indel similarity and the token set ratio. The former is the minimum number of insertions and deletions to convert one string into the other. The latter is the number of tokens in the shorter string contained in the longer string divided by the number of tokens in the longer string. We validated the use of the token ratio by the MAP on the validation dataset (cf. Section 3). We tested matching song title and performer concatenated and using solely the song title of the query item. For the candidate item, we experimented with only the video title and with the latter combined with the other attributes concatenated by space. The best configuration was simply matching the song title to the video title (cf. *Simple* input in Figure 1). We also experimented with the snowball stemmer from NLTK (Bird et al., 2009) for all attributes but it did not improve the results.

S-BERT The model S-BERT (Reimers and Gurevych, 2019) addresses the problem of quadratic complexity of language models (LMs) like BERT (Devlin et al., 2018) when processing pairs of sentences. Hence, it was used as promising method for ER (Li et al., 2021a; Paganelli et al., 2023). The model learns to encode sentences into embedding vectors which can be compared using similarity measures such as Cosine similarity. We fine-tune a multilingual approach of S-BERT (Reimers and Gurevych, 2020) which encodes text sequences into 384-dimensional vectors. Like for fuzzy matching, we use the *Simple* input because it performed better than the other attribute combinations. We apply a similar training procedure like in recent CSI approaches: we use triplet loss with a margin of 0.3 and apply online hard triplet mining (Xuan et al., 2020) where the hardest triplets in the batch of 16 items with 4 random works represented by 4 items are used for training updates. We select the best model after 10 epochs measured in MAP on the validation dataset.

Ditto Other state-of-the-art approaches rely on contextualized embeddings provided by pretrained LMs to predict the matching confidence for a given entity pair. In theory, this can improve the ER task since the context is not only considered for tokens in one entity but across both entities. Thus, we experimented with Ditto (Li et al., 2020), HierGAT (Yao et al., 2022), and r-SupCon (Peeters and Bizer, 2022). We found that Ditto was both – better performing and faster in inference. We therefore select Ditto for our experiments. The model computes a binary matching confidence based on entity pairs encoded by LMs such as RoBERTa (Liu et al., 2019). Due to the quadratic complexity during inference, we use S-BERT as blocker. We adopt the top-k blocking strategy suggested by the authors (Li et al., 2021b) where the top-k most similar candidate items per query item are passed to Ditto and the remaining pairs are predicted by the blocker.

As underlying LM, we chose RoBERTa since it was shown to achieve high performance in the ER task (Li et al., 2020; Peeters and Bizer, 2022; Peeters et al., 2023; Yao et al., 2022) and multilingual variant of BERT (mBERT) (Devlin et al., 2018). First, we experimented with the same attribute combinations we use for fuzzy matching and S-BERT. We observed that Ditto works better

when items are represented with the same attributes on the query and candidate side. Hence, we use a *rich input* (cf. Figure 1). As inputs for the LMs, we concatenate the names and values of attributes of the query and candidate item: Each attribute is represented by a [COL] token followed by its name (e.g., *title* for the song title) and a [VAL] token followed by its value (e.g., *Yesterday*). Since the actual song attributes of the items on the candidate side are not known, we mask the respective tokens of those using a [MASK] token. We fine-tune with a batch size of 32 and a learning rate of 1e-05 and a sequence length of 256 tokens. We select the best performing model after 15 epochs. We fine-tune Ditto with a dataset of pairs (cf. Section 3) and measure the performance on the validation dataset in F1.

2.2 Combining ER with CSI

We form multi-modal ensembles each combining one fine-tuned ER model with a trained CSI model (ER-CSI ensembles). We use two pre-trained CSI models: CQTNet (Yu et al., 2020) and CoverHunter (Liu et al., 2023). Both models encode items into vectors and represent musical similarity using Cosine similarity. The former uses convolutional neural networks to learn 300-dimensional vector representations. CoverHunter uses conformer neural networks (Gulati et al., 2020) and an attention mechanism for temporal pooling (Okabe et al., 2018) to learn 128-dimensional vector representations. For each ER-CSI ensemble we train the ranking model LambdaMART (Wu et al., 2010) using the pairwise similarities as input features. We use (mean average precision) MAP objective function and consider the top 50 feature interactions similar to (Lucchese et al., 2022).

3 Experimental Setup

Our experiments aim to evaluate a) whether ensembles of ER and CSI models outperform CSI models and b) whether ER models are robust against hard negatives and song title variations (e.g., translations of song titles or parodies).[3] We report two evaluation metrics suggested by MIREX:[4] MAP and mean rank of the first relevant item (MR1).

We use subsets of two popular CSI datasets: SHS100K (Xu et al., 2018) and DaTacos (Yesiler et al., 2019), which are popular in CSI research. Both datasets contain items which are songs represented by metadata attributes, a YouTube identifier, and a work identifier. We only retain items with available videos and denote the resulting datasets as V-SHS and V-DaT respectively (cf. Table 1). We retrieve YouTube metadata for all the videos using *YouTube Search Python* and extract the audio features as described by the authors of CoverHunter (Liu et al., 2023) and CQTNet (Yu et al., 2020).

Dataset	Split	Works	Avg.	FM	DI	SB
V-SHS-F	Train	1,501	1.33	✗	✓	✓
V-SHS-V	Valid	1,827	4.73	✓	✗	✓
V-SHS-P	Valid	1,224	1,63	✗	✓	✗
V-SHS-T	Test	1,679	5.25	✓	(✓)	✓
-Unique	Test	852	2.75	✓	✓	✓
-Noise	Test	12	4.15	✓	✓	✓
V-DaT	Test	2,784	4.92	✓	(✓)	✓

Table 1: Dataset statistics: the number of works (Works) with average number of items (Avg.) and usage for Fuzzy Matching (FM), Ditto (DI), or S-BERT (SB). (✓) denotes partial use (after blocking).

Training and Validation Subsets We fine-tune S-BERT and Ditto and train LambdaMART using a random sample of items from the V-SHS100K training set with 1,000 positive pairs and 6,000 negative pairs similar to datasets by Konda et al. (2016). Similarly, we create a pair-wise validation dataset used to select the best model checkpoint of Ditto. The full overview of datasets used is given in Table 1.

Test Subsets Addressing a), we select the *V-DaT* and *V-SHS-T* which are subsets with available videos of datasets typically used in CSI evaluation. For b), the robustness study, we create additional test datasets. First, we only retain one item per work and song title (dropping multiple items with the same song title per work). This subset based on *V-SHS-T* is denoted by *V-SHS-T-Unique*.

Lastly, we aim to evaluate the ER models' robustness to hard negatives. On YouTube, this can be expected in cases where either the song title is the same for different works or when the words in the song title are used in a different context (e.g., the song title "Hush" occurring in the sentence "Relaxing Hush Sounds"). We focus on the latter problem which is particularly challenging for song titles with one word, because these words are more likely to occur in different contexts. We create

[3] An example for a parody title is "Bye, Bye Johnny" by The Rattles covering "Johnny B. Goode"

[4] cf. https://www.music-ir.org/mirex/wiki/2021:Audio_Cover_Song_Identification

Song Title	Utterances in generated video title
Yesterday	**Yesterday**'s Kitchen: Old Recipes
Hush	Relaxing **Hush** Sounds
Time	Mastering **Time** Management

Table 2: Examples of one-word song titles and video titles of generated hard negatives with ChatGPT 3.5 in the SHS100K-T-Noise dataset.

a subset of *V-SHS-T* containing only items with one-word song titles. For each of the works in the dataset we instruct ChatGPT 3.5 (OpenAI Inc.) to generate video metadata containing the respective word.[5] We show some examples in Table 2. The resulting dataset *V-SHS-T-Noise* contains 12 works each with 5 generated video titles (hard negatives) and on average around 4 items of the work.

4 Results

In Table 3 we report the results of experiment a). Generally, we observe a strong improvement in MR1 when comparing ER-CSI ensembles with CSI. Improvements in MAP are evident but smaller for ensembles with CoverHunter. Still, gains up to +9% in MAP and up to -13.45 ranks in MR1 are achieved on V-DaTacos. The ensembles with Fuzzy Matching have relatively small performance gains and do not achieve a higher MAP than CoverHunter.

The highest results in MAP are achieved with S-BERT and Ditto with S-BERT as blocker and RoBERTa as underlying LM. Considering the computational overhead for using Ditto, even with blocking and $k = 100$, makes its use for the task questionable. However, Table 4 shows that combining Ditto with S-BERT can stabilize robustness in some cases. Even though the latter achieves higher MAP on the -Unique subsets, combining it with Ditto (M) yields +5% and +8% in MAP on those. Apparently, this might be due to better rankings of Ditto at the earlier ranks indicated by MR1.

5 Limitations and Conclusion

In this paper, we implemented ER approaches as means to support the task of cover song identification in online videos on the example of the online video platform YouTube. We showed that simple fuzzy matching can partly help to increase model performances. Better results were achieved by using S-BERT. Additionally, using Ditto appears to be adequate only in some cases such as song title variations. However, the robustness of ER models is generally harmed by hard negatives which potentially do not refer to music.

Lastly, we outline limitations of this study. Our selected text input structure for S-BERT and fuzzy matching can only detect the song title in the video titles. While this might be sufficient for our utilized datasets, other videos which contain the title only in the description or keywords are not uncovered. Secondly, song titles that are completely different than the reference (e.g., parodies or medleys) cannot be detected by ER models. Hence, we see these models as supporting tool in OCSI rather than independent approaches. In future research, we aim to leverage more recent large LMs which are starting to get used for ER (Peeters and Bizer, 2024).

[5]The prompt was: *Consider the following a list of words. Generate a meaningful video title for each of these words.*

	V-SHS-T		**V-DaT**	
	MAP	MR1	MAP	MR1
CQTNet	0.71	47.40	0.74	10.74
-Fuzzy	0.75	14.92	0.80	4.16
-S-BERT	**0.85**	**12.14**	0.91	3.06
-SB+Ditto (R)	**0.85**	16.29	**0.92**	**3.03**
-SB+Ditto (M)	0.83	20.62	0.90	3.49
CoverHunter	0.92	12.60	0.84	15.71
-Fuzzy	0.90	4.47	0.84	5.57
-S-BERT	**0.93**	**3.58**	**0.93**	3.00
-SB+Ditto (R)	**0.93**	5.41	**0.93**	**2.26**
-SB+Ditto (M)	0.92	7.06	0.91	2.70

Table 3: Experiment a): Performances of ER-CSI ensembles against CSI models. SB+Ditto denotes Ditto with S-BERT as blocker; (R) stands for RoBERTA and (M) for mBERT.

	-Noise		**SHS-Uniq.**	
	MAP	MR1	MAP	MR1
Fuzzy	0.38	5.03	0.37	189.78
S-BERT	0.53	4.15	0.55	138.04
Ditto (R)	0.43	**2.97**	0.37	114.94
Ditto (M)	0.49	4.21	0.46	**100.26**
SB+Ditto (R)	**0.56**	**2.97**	0.51	183.22
SB+Ditto (M)	0.44	7.31	**0.60**	252.31

Table 4: Experiment b): Results of ER approaches on the V-SHS-T-Noise (-Noise), V-SHS-T-Unique (S-Uniq.). Due to the smaller dataset size, we set $k = 10$ for -Noise.

References

Mathilde Abrassart and Guillaume Doras. 2022. And what if two musical versions don't share melody, harmony, rhythm, or lyrics ? In *International Society for Music Information Retrieval Conference*.

Swati Agrawal and Ashish Sureka. Copyright Infringement Detection of Music Videos on YouTube by Mining Video and Uploader Meta-data. In *Big Data Analytics*, Lecture Notes in Computer Science, pages 48–67. Springer International Publishing.

Massimo Airoldi, Davide Beraldo, and Alessandro Gandini. 2016. Follow the algorithm: An exploratory investigation of music on youtube. *Poetics*, 57:1–13.

Max Bachmann. 2021. maxbachmann/rapidfuzz: Release 1.8. 0.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Xingjian Du, Zijie Wang, Xia Liang, Huidong Liang, Bilei Zhu, and Zejun Ma. 2023. Bytecover3: Accurate cover song identification on short queries. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Shichao Hu, Bin Zhang, Jinhong Lu, Yiliang Jiang, Wucheng Wang, Lingcheng Kong, Weifeng Zhao, and Tao Jiang. 2022. WideResNet with Joint Representation Learning and Data Augmentation for Cover Song Identification. In *Proc. Interspeech 2022*, pages 4187–4191.

Pradap Konda, Sanjib Das, Paul Suganthan G. C., AnHai Doan, Adel Ardalan, Jeffrey R. Ballard, Han Li, Fatemah Panahi, Haojun Zhang, Jeff Naughton, Shishir Prasad, Ganesh Krishnan, Rohit Deep, and Vijay Raghavendra. 2016. Magellan: Toward building entity matching management systems over data science stacks. *Proc. VLDB Endow.*, 9(13):1581–1584.

Bing Li, Yukai Miao, Yaoshu Wang, Yifang Sun, and Wei Wang. 2021a. Improving the efficiency and effectiveness for bert-based entity resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13226–13233.

Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. 2020. Deep entity matching with pre-trained language models. *Proceedings of the VLDB Endowment*, 14(1):50–60.

Yuliang Li, Jinfeng Li, Yoshihiko Suhara, Jin Wang, Wataru Hirota, and Wang-Chiew Tan. 2021b. Deep entity matching: Challenges and opportunities. *Journal of Data and Information Quality (JDIQ)*, 13(1):1–17.

Lassi A. Liikkanen and Antti Salovaara. 2015. Music on youtube: User engagement with traditional, user-appropriated and derivative videos. *Computers in Human Behavior*, 50:108–124.

Feng Liu, Deyi Tuo, Yinan Xu, and Xintong Han. 2023. Coverhunter: Cover song identification with refined attention and alignments. *Preprint*, arXiv:2306.09025.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, and Alberto Veneri. 2022. Ilmart: Interpretable ranking with constrained lambdamart. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2255–2259.

Alberto Montero and Jorge Mora-Fernandez. 2020. Digital culture in youtube categories and interfaces: User experience and social interactions of the most popular videos and channels. In *International Conference on Human-Computer Interaction*, pages 383–401. Springer.

Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda. 2018. Attentive statistics pooling for deep speaker embedding. In *Interspeech 2018*, interspeech$_2$018.$ISCA$.

OpenAI Inc. ChatGPT.

Matteo Paganelli, Donato Tiano, and Francesco Guerra. 2023. A multi-facet analysis of bert-based entity matching models. *The VLDB Journal*, pages 1–26.

Ralph Peeters and Christian Bizer. 2022. Supervised contrastive learning for product matching. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 248–251, New York, NY, USA. Association for Computing Machinery.

Ralph Peeters and Christian Bizer. 2024. Entity matching using large language models. *Preprint*, arXiv:2310.11244.

Ralph Peeters, Reng Chiz Der, and Christian Bizer. 2023. Wdc products: A multi-dimensional entity matching benchmark. *arXiv preprint arXiv:2301.09521*.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Jordan B. L. Smith, Masahiro Hamasaki, and Masataka Goto. 2017. Classifying derivative works with search, text, audio and video features. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1422–1427.

Qiang Wu, Christopher J. C. Burges, Krysta M. Svore, and Jianfeng Gao. 2010. Adapting boosting for information retrieval measures. 13(3):254–270.

Xiaoshuo Xu, Xiaoou Chen, and Deshun Yang. 2018. Key-invariant convolutional neural network toward efficient cover song identification. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Hong Xuan, Abby Stylianou, Xiaotong Liu, and Robert Pless. 2020. Hard negative examples are hard, but useful. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 126–142. Springer.

Dezhong Yao, Yuhong Gu, Gao Cong, Hai Jin, and Xinqiao Lv. 2022. Entity resolution with hierarchical graph attention networks. In *Proceedings of the 2022 International Conference on Management of Data*, SIGMOD '22, page 429–442, New York, NY, USA. Association for Computing Machinery.

Furkan Yesiler, Chris Tralie, Albin Correya, Diego F. Silva, Philip Tovstogan, Emilia Gómez, and Xavier Serra. 2019. Da-TACOS: A dataset for cover song identification and understanding. In *Proc. of the 20th Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, pages 327–334, Delft, The Netherlands.

Zhesong Yu, Xiaoshuo Xu, Xiaoou Chen, and Deshun Yang. 2020. Learning a representation for cover song identification using convolutional neural network. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 541–545.

Can Impressions of Music be Extracted from Thumbnail Images?

Takashi Harada
The University of Tokyo

Takehiro Motomitsu
Hokkaido University

Katsuhiko Hayashi
The University of Tokyo

Yusuke Sakai
NAIST

Hidetaka Kamigaito
NAIST

katsuhiko-hayashi@g.ecc.u-tokyo.ac.jp

Abstract

In recent years, there has been a notable increase in research on machine learning models for music retrieval and generation systems that are capable of taking natural language sentences as inputs. However, there is a scarcity of large-scale publicly available datasets, consisting of music data and their corresponding natural language descriptions known as music captions. In particular, non-musical information such as suitable situations for listening to a track and the emotions elicited upon listening is crucial for describing music. This type of information is underrepresented in existing music caption datasets due to the challenges associated with extracting it directly from music data. To address this issue, we propose a method for generating music caption data that incorporates non-musical aspects inferred from music thumbnail images, and validated the effectiveness of our approach through human evaluations. Additionally, we created a dataset with approximately 360,000 captions containing non-musical aspects. Leveraging this dataset, we trained a music retrieval model and demonstrated its effectiveness in music retrieval tasks through evaluation.

1 Introduction

The enjoyment of music is a highly personal experience, and the ways in which individuals find pleasure in it vary greatly. One method of enjoying music involves listening to tracks that best suit one's current mood or situations, such as during events like birthdays or in accordance with the season. Furthermore, individuals often find pleasure in music through discovering songs that match their preferred genres or instruments, and even in creating new music by utilizing the characteristics of existing music. As our interaction with music becomes more sophisticated, the development of music retrieval and generation systems that are customized to individual tastes and preferences is increasingly important.

	Musical aspects	**Non-musical aspects**
Perspective	- genre - instruments - tempo	- suitable situations - seasons, times of day - evoked emotions
Music caption	"Modern jazz played with a triumphal trumpet"	"Calm song ideal for listening on a night with a visible starry sky"

Table 1: Examples of musical and non-musical aspects.

In the domain of music retrieval, studies such as MusCALL (Manco et al., 2022a) have investigated the potential for efficient retrieval of music using natural language sentences as inputs, employing contrastive learning methodologies (Le-Khac et al., 2020). Efforts similar to those of MuLan (Huang et al., 2022) are currently being pursued in the field. In the realm of music generation, research initiatives such as MusicLM (Agostinelli et al., 2023), MUSICGEN (Copet et al., 2023), Mubert (Mubert-Inc, 2022), and Riffusion (Forsgren and Martiros, 2022) are integrating elements from systems like MuLan to produce high-quality music from textual descriptions. These music retrieval and generation models are dependent on natural language descriptive texts, commonly known as music captions, which contain information pertinent to music.

However, a notable challenge emerges, as the bulk of music caption data utilized in these models and systems is not accessible to the public, largely due to copyright constraints. Furthermore, because existing descriptive texts primarily focus on musical information, they hinder users without extensive musical knowledge from efficiently conducting music retrieval and generation that is based solely on musical elements. Consequently, there is an urgent need to develop and disseminate music caption data that includes a variety of elements, both musical and non-musical aspects, to the public.

To ensure the diversity of music caption data,

it is imperative to include descriptions from two distinct perspectives: musical and non-musical aspects, as delineated in Table 1. The musical aspects encompass information like genre and tempo, which can be extracted from music data through music information processing, as exemplified by initiatives such as LP-MusicCaps (Doh et al., 2023) and others (Liu et al., 2023c; Manco et al., 2021). In contrast, non-musical aspects include individual impressions and emotional responses associated with music, encompassing elements such as suitable situations for listening to a track, associated seasons and times of day. The direct extraction of such information from music data poses significant challenges. Consequently, there is a notable deficiency in music caption data that includes non-musical aspect descriptions, and efforts toward their automatic generation are markedly limited.

To address this issue, we suggest focusing on the thumbnail images associated with music clips on platforms such as YouTube to extract information pertaining to non-musical aspects. Generally, thumbnail images significantly influence user engagement and click-through rates for content. In the context of music clips, thumbnail images serve as a succinct visual representation of the music's impression, enabling users to decide at a glance whether it aligns with their music preferences. In this research, we introduce a methodology for the automatic generation of music caption data, enriched with non-musical aspects derived from these thumbnail images. The validation results demonstrate that our proposed method is capable of accurately generating music captions that express non-musical aspect information, outperforming baseline methods. Furthermore, we have made public a dataset (`https://github.com/hu-kvl/llava_music_caption`) containing approximately 360,000 music captions developed using our approach. The utility of this dataset was also assessed within a music retrieval model, further substantiating its effectiveness. The details regarding the dataset creation and the evaluation of the music retrieval model are provided in the Appendices.

2 Related Work

To address the scarcity of music caption datasets, the LP-MusicCaps (LLM-Based Pseudo Music Captioning) initiatively utilized GPT-3.5 Turbo (OpenAI, 2022), a Large Language Model (LLM), for generating music captions across various tasks employing a comprehensive music tag dataset (Law et al., 2009; Bertin-Mahieux et al., 2011). The dataset produced, encompassing approximately 2.2 million captions and 500,000 audio clips paired with tags, has been made publicly available. It facilitates the automated generation of music captions with an enriched vocabulary while ensuring alignment with tags and grammatical precision. Notably, these methods utilize LLMs to focus on generating music captions that incorporate musical elements derived from tags, in contrast to our approach of enriching captions with non-musical aspects.

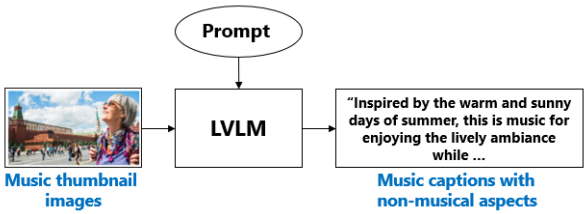


Figure 1: Overview of the music captioning method using thumbnail images.

While our study focuses on the automatic generation of music captions enriched with non-musical aspects using thumbnail images, there exists related research that leverages image intermediary representations for generating lyrics (Watanabe and Goto, 2023). This research tackles the challenge of assisting users in conveying appropriate messages and words in lyric creation.

Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) uses a dual-encoder architecture to embed images and text into a shared latent space. CLIP has been influential in its demonstration of extracting highly generic visual representations from natural language, and MusCALL (Manco et al., 2022a) also adopts this dual-encoder contrastive learning approach pioneered by CLIP. MuLan (Huang et al., 2022) represents a similar endeavor in this field. Additionally, CLAP (Wu et al., 2023) is a model that applies CLIP's methodology to the audio domain. A cross-modal retrieval method between music and images has been also investigated in (Nakatsuka et al., 2023). In this study, we train a music retrieval model using pairs of music captions and music data created through the proposed method, and assess its effectiveness in music retrieval tasks where input queries are written in natural language sentences.

Evaluation Metric	Details of Evaluation Metric	Score
Positive	Assuming a music generation model that generates music from the provided text, this text contains and accurately expresses the corresponding non-musical aspects.	2
Neutral	Assuming a music generation model that generates music from the provided text, this text contains the corresponding non-musical aspects, but their expression is not accurate.	1
Negative	This text does not contain the corresponding non-musical aspects.	0

Table 2: Metrics for the human evaluation.

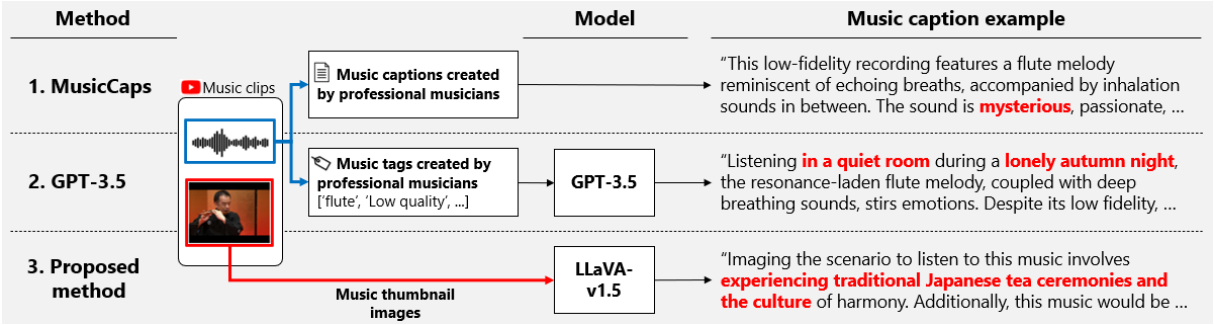


Figure 2: Details of the proposed music captioning method and the methods used for comparison.

3 Proposed Method

This study is centered around thumbnail images linked to music clips on platforms like YouTube. Figure 1 provides an overview of our proposed methodology. In our approach, we first input the thumbnail image into a Large Vision-Language Model (LVLM), a large-scale language model capable of processing images. Following this, the LVLM, adapted using a meticulously crafted prompt, generates a music caption. This procedure facilitates the automatic generation of music captions that incorporate non-musical elements.

To ensure that the features of the thumbnail images are not included in the captions generated by the LVLM, we developed a strategic prompting method. Specifically, we guide the model to initially describe the features of the image itself, such as the content depicted in the thumbnail or its overall mood, in the initial segment of the generated text. In the subsequent part, we prompt the model to articulate non-musical aspects that are evoked by these features or the overall mood.

Moreover, by directing the LVLM to distinctly express the features of the thumbnail image in the first portion of the generated text, we successfully separate the description of thumbnail image features from the music caption generation. This not only achieves effective compartmentalization but also leverages techniques like the Chain of Thought (Wei et al., 2022), resulting in the production of high-quality music captions enriched with diverse non-musical elements.

4 Human Evaluation

4.1 Experimental Setup

For the generation of captions to be evaluated, we utilized the MusicCaps (Agostinelli et al., 2023) dataset[1]. MusicCaps primarily includes descriptions of music information from YouTube music clips, including corresponding music captions and music tags created by professional musicians. In our study, we selected 50 songs from MusicCaps, ensuring a balanced distribution across genres, for evaluation. For each song in the evaluation, three types of captions were prepared:

1. Manually created captions originally assigned to the dataset (MusicCaps).
2. Captions automatically generated using a LLM in accordance with the LP-MusicCaps (Doh et al., 2023) study approach, employing music tags (GPT-3.5).
3. Captions automatically generated from thumbnail images using the LVLM (Proposed method).

The distinct caption generation processes for each method are depicted in Figure 2. For GPT-3.5, we employed the GPT-3.5 Turbo language model, instructing it to describe non-musical aspects based on music tags and those tags them-

[1] https://huggingface.co/datasets/google/MusicCaps

	MusicCaps	GPT-3.5	**Proposed Method**
Situation	39.0	46.5	**75.5**
Time/Season	6.0	35.0	**72.0**
Emotion	36.5	**89.0**	83.0
Total	81.5	170.5	**230.5**
All 2s count	1.5	13.0	**23.5**

Table 3: Human evaluation scores and the count of All 2s.

selves. For the proposed method, we selected the open-source LLaVA-v1.5 (Liu et al., 2023b,a)[2] as the LVLM and tuned the prompt to align with our proposed method. The average lengths of the captions were 148.7 characters for MusicCaps, 146.0 characters for GPT-3.5, and 135.8 characters for the proposed method.

To evaluate each set of captions, a human evaluation was conducted. For the proposed method, only sections of the generated results related to non-musical aspects, located in the latter part of the output, were extracted for evaluation.

4.2 Evaluation Procedure and Metrics

For the human evaluation component, we enlisted two adult Japanese male speakers as evaluators. These evaluators listened to each music clip for approximately one minute before reviewing the three captions generated by each method. They then rendered absolute evaluations using a three-point scale (positive, neutral, negative) based on three perspectives (we referred to (Manco et al., 2022b)): suitable situations for listening to the track (Situation), associated times of day or seasons (Time/Season), and the emotions evoked as a result of listening (Emotion), as delineated in Table 2. To circumvent order bias, the captions were presented in a randomized sequence.

To facilitate the aggregation of results, the evaluators' three-tier assessments were converted into scores, as illustrated in Table 2. The final results reflect the average scores across evaluators, providing a cumulative score for the full dataset of 50 entries for each perspective. Furthermore, the aggregate values for each perspective and the average count of tracks where all three perspectives achieved a perfect score, designated as the "All 2s count," are also reported. It is predicated on the premise that captions with higher scores and greater counts not only contain non-musical aspects but also exhibit expressiveness.

[2]We performed 4-bit inference on LLaVA-v1.5 13B.

4.3 Results and Discussion

The evaluation results are presented in Table 3, and actual music caption examples are shown in Figure 2, using the case of MusicCaps with "YouTube Video ID (ytID):0u5-WiBKam8" as an illustration.

Focusing on the aggregate scores across three perspectives, our proposed method achieved the highest evaluation outcomes. When compared to GPT-3.5, our method surpassed it by a total of 60 points in aggregate scores. Based on this result, it can be inferred that music captions, including non-musical aspects, can potentially be generated using information extracted from thumbnail images alone, without relying on musical or non-musical tag information created by professional musicians. Moreover, in every perspective, the scores of both our method and GPT-3.5 exceeded those of manually created captions by humans, MusicCaps. This suggests that when evaluating music captions based on three perspectives, LVLMs and LLMs can potentially be used to efficiently generate music captions that include non-musical aspects.

Finally, with a focus on the actual music caption examples, Figure 2 displays instances where non-musical aspects are expressed, written in red letters. Captions generated by our method are detailed and expressive, providing a rich representation of non-musical aspects. In contrast, captions by MusicCaps and GPT-3.5 include fewer descriptions of non-musical aspects and are more abstract.

5 Conclusion

In this study, we examined the importance of non-musical aspects, such as suitable situations for listening, times of day or seasons, and emotions evoked by listening. Specifically, we addressed the issue of the shortage of music caption data that includes non-musical aspect information, which is essential for constructing music retrieval and generation models. To address this issue, we proposed a method for generating music captions by leveraging thumbnail images. The effectiveness of the proposed method was validated through human evaluations. It has been reported that using the latest LVLMs to generate descriptions related to artworks has been successful (Hayashi et al., 2024; Saito et al., 2024; Ozaki et al., 2024), and we plan to update our experiments in accordance with the further advancements of LVLMs.

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. arXiv:2301.11325. *Preprint*, arXiv:2301.11325.

Thierry Bertin-Mahieux, Daniel Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, pages 591–596.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. arXiv:2306.05284. *Preprint*, arXiv:2306.05284.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. 2017. Fma: A dataset for music analysis. arXiv:1612.01840. *Preprint*, arXiv:1612.01840.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023. LP-MusicCaps: LLM-based pseudo music captioning. arXiv:2307.16372. *Preprint*, arXiv:2307.16372.

S. Forsgren and H Martiros. 2022. Riffusion. `https://riffusion.com/about`. Accessed: January 16, 2024.

Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. 2003. Rwc music database: Music genre database and musical instrument sound database.

Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. Towards artwork explanation in large-scale vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 705–729, Bangkok, Thailand. Association for Computational Linguistics.

Kun He, Yan Wang, and John Hopcroft. 2016. A powerful generative model using random weights for the deep image representation. arXiv:1606.04801. *Preprint*, arXiv:1606.04801.

Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P. W. Ellis. 2022. MuLan: A joint embedding of music audio and natural language. arXiv:2208.12415. *Preprint*, arXiv:2208.12415.

Edith Law, Kris West, Michael I. Mandel, Mert Bay, and J. S. Downie. 2009. Evaluation of algorithms using games: The case of music tagging. In *International Society for Music Information Retrieval Conference*.

Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. 2020. Contrastive representation learning: A framework and review. *Ieee Access*, 8:193907–193934.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. arXiv:2310.03744. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. arXiv:2304.08485. *Preprint*, arXiv:2304.08485.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2023c. Music understanding LLaMA: Advancing text-to-music generation with question answering and captioning. arXiv:2308.11276. *Preprint*, arXiv:2308.11276.

Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. 2021. MusCaps: Generating captions for music audio. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. 2022a. Contrastive audio-language learning for music. arXiv:2208.12208. *Preprint*, arXiv:2208.12208.

Ilaria Manco, Benno Weck, Philip Tovstogan, Minz Won, and Dmitry Bogdanov. 2022b. Song describer: a platform for collecting textual descriptions of music recordings. In *Extended Abstracts for the Late-Breaking Demo Session of the 23rd Int. Society for Music Information Retrieval Conf*, Bengaluru, India.

Mubert-Inc. 2022. Mubert. `https://mubert.com,https://github.com/MubertAI/Mubert-Text-to-Music`. Accessed: January 16, 2024.

Takayuki Nakatsuka, Masahiro Hamasaki, and Masataka Goto. 2023. Content-based music-image retrieval using self-and cross-modal feature embedding memory. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2174–2184.

OpenAI. 2022. `https://openai.com/chatgpt`. Accessed: November 26, 2023.

Shintaro Ozaki, Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. Towards cross-lingual explanation of artwork in large-scale vision language models. *arXiv preprint arXiv:2409.01584*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Shigeki Saito, Kazuki Hayashi, Yusuke Ide, Yusuke Sakai, Kazuma Onishi, Toma Suzuki, Seiji Gobara, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. Evaluating image review ability of vision language models. *arXiv preprint arXiv:2402.12121*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Kento Watanabe and Masataka Goto. 2023. Text-to-lyrics generation with image-based semantics and reduced risk of plagiarism. In *Ismir 2023 Hybrid Conference*.

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

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 *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

A Appendices

A.1 Open Music Caption Dataset

In this research, we employed the proposed method to create an open music caption dataset comprising approximately 360,000 pairs of music clips and their associated non-musical aspect captions. For this dataset, 15 music genres were selected based on their common usage in music databases (Goto et al., 2003), datasets (Defferrard et al., 2017), and streaming services. Table 4 provides a detailed breakdown of each genre along with the corresponding number of data entries. The music clips used were sourced by searching for the respective genre names and using the tracks found in playlists, with no intentional filtering applied. However, some clips that could not be acquired due to reasons such as privacy or age restrictions were excluded.

The construction of the training dataset entailed utilizing the proposed method to input thumbnail images of YouTube music clips and Prompt1 (see Figure 3) into the LLaVA model. Prompt1 is designed to facilitate the generation of content across five sections, encompassing the description of the mood and features of the thumbnail image, the preferred listening situation, scenario, and setting, the ideal times of day and seasons for listening, the emotions experienced while listening to the music, and a summarizing sentence that encapsulates the aspects detailed in sections 2, 3, and 4. The structured prompts were devised to distinctly separate the description of image features from the caption generation process, thereby enhancing the efficacy of the Chain of Thought approach.

The resulting training dataset includes fields such as "YouTubeID," "URL," "Genre," "Caption," and "Sentence" for each music clip. The "Sentence" field contains the entire content of Prompt1, enabling its potential use as a comprehensive, non-musical aspect-inclusive music caption.

Genre	Train	Test (All 2s)	Total
house	49,942	80 (49)	50,022
edm	41,406	80 (53)	41,486
classic	32,695	80 (55)	32,775
chill	31,888	80 (62)	31,968
lofi	27,084	80 (62)	27,164
nightcore	24,717	80 (46)	24,797
anime	24,664	80 (55)	24,744
pop	24,658	80 (53)	24,738
rock	24,425	80 (37)	24,505
instrumental	23,483	80 (59)	23,563
tropical house	21,655	80 (56)	21,735
jazz	12,676	80 (51)	12,756
r&b	8,258	80 (60)	8,338
hiphop	7,812	80 (49)	7,892
bigroom	5,542	80 (43)	5,622
Total	360,905	1,200 (790)	362,105

Table 4: Number of data samples by genre.

This image represents a thumbnail of a music clip. Please provide a response to the following sections in this format.
1. Please describe the mood and features depicted in the image freely.
2. Describe the ideal listening situation, scenario, and setting for music.
3. Describe the preferred time of day and seasons ideal for listening to music.
4. Describe the emotions felt when listening to the music for music.
5. Summarize the content from Sections 2, 3, and 4 into a single sentence, ensuring it absolutely includes the desired listening situation, time or season for listening, and the emotional when listening to the music within one sentence.

Figure 3: Prompt 1: prompt to generate music captions from thumbnail images.

A.2 Evaluation Dataset

For the creation of the evaluation dataset, our objective was to generate ground truth data that aligns with human perceptions and sensitivities towards music. From the training dataset, we randomly selected 80 music clips for each of the 15 genres, resulting in a total of 1,200 clips for evaluation. These clips were assessed by a musically knowledgeable adult using a three-point scale based on the perspectives detailed in Section 4, following approximately a minute of listening to each clip. The evaluated aspects included the ideal listening situation, the most suitable times of day or seasons for listening, and the emotions experienced during listening.

The evaluation dataset comprises fields such as "YouTubeID," "URL," "Genre," and "Caption," as well as "Situation," "Time/Season," and "Emotion" for each music clip. The latter three fields encapsulate the scores obtained from the human evaluations, corresponding to the three outlined perspectives. It is crucial to note that both the training and evaluation datasets are entirely in English. Table 4 details the breakdown and data count for each genre, including the number of music clips that received a "All 2s."

A.3 Validation with music retrieval Model

To evaluate the appropriateness of the newly created dataset for general user retrieval, we conducted an assessment by training a music retrieval model on the music caption dataset. The efficacy of the model is gauged based on its proficiency in accurately retrieving relevant items from the dataset.

A.3.1 Experimental Setup

To validate the effectiveness of the created dataset, we utilized the MusCALL music retrieval model (Manco et al., 2022a). MusCALL is a cross-modal contrastive learning model that facilitates both text-based music retrieval and music-based natural language retrieval. The architecture of the model includes a Transformer (Vaswani et al., 2017; Radford et al., 2019) for text encoding and a ResNet-50 (He et al., 2016), equipped to handle audio input, for audio encoding. The embeddings generated by these encoders are utilized to calculate cross-entropy loss, aimed at minimizing the cumulative losses of text and audio embeddings. The model's configurations, such as hyperparameters, are aligned with those established in MusCALL.

Genre	R@1↑	R@5↑	R@10↑	MedR↓
house	12.2	38.8	**59.2**	**8**
anime	5.5	27.3	**56.4**	**9**
instrumental	16.9	37.3	**57.6**	**9**
jazz	7.8	35.3	51.0	10
classic	10.9	36.4	50.9	10
pop	5.7	20.8	47.2	11
rock	10.8	35.1	46.0	11
chill	6.5	24.2	37.1	14
nightcore	2.2	19.6	37.0	14
tropical house	10.7	23.2	41.1	14
hip hop	2.0	28.6	42.9	15
big room	0.0	9.3	25.6	17
edm	7.5	17.0	32.1	17
lofi	12.9	30.6	38.7	17
r&b	3.3	20.0	28.3	23
Average	7.7	26.9	43.4	13.3

Table 5: Results of text to audio cross-modal retrieval task.

A.3.2 Dataset

For the training and evaluation of the MusCALL model, three distinct datasets - train, validation, and test - are employed. To construct the validation dataset, we randomly selected 30,000 entries from the training data, as delineated in Section A.1. Furthermore, 790 entries that are all categorized as "All 2s" from the 1,200 human-evaluated music clips were utilized as the test dataset. A detailed breakdown of the data utilization is provided in Table 4.

It is crucial to emphasize that the audio data used for this validation consists of 30-second clips extracted from YouTube music clips. These clips are accompanied by captions that are derived from the non-musical aspects indicated in the dataset's "Caption" column.

A.3.3 Evaluation Metrics

This experiment fundamentally represents a cross-modal retrieval task, tasked with evaluating the model's proficiency in retrieving items across different modalities. To assess the model's retrieval accuracy, we employ standard cross-modal retrieval metrics, namely Recall at K (R@K) and Median Rank (MedR). Specifically, we set K = {1, 5, 10}, which entails measuring the percentage of queries where the correct pair is ranked within the top k and the median rank of the correct pair, respectively, across various genres.

A.3.4 Results and Discussion

The outcome of the cross-modal retrieval tasks, namely retrieving music from text, is delineated in Table 5 for each genre. In genres such as house,

Ranking	YouTubeID	Thumbnail	Title	Evaluation
1	POHHAe0eTfA		ALEJANDRO - Black Panther	Good
2	W3H_8l6dvZs		Crypto - Faded (feat. Constance)	Fair
3	kFDfvMr1Pd4		JOXION - 094 [Arcade Release]	Excellent
4	N4Up97ZOv0g		Sora - Changes	Good
5	LL1owMMYP78		Black Gypsies - Kira (Original Mix)	Poor
6	AWXsZYhlmwg		Free My Body	Excellent
7	8xn5gH3XUkQ		New Blood & RCOP - Murda Sound	Good
8	JY97C7EztAg		S'hustryi Beats - Anti-Covid19	Poor
9	q3ugvrgOri4		Close Your Eyes (Extended Mix)	Good

Table 6: Retrieval results and evaluations for the query "Energetic songs for a late-night drive."

anime, instrumental, jazz, classic, pop, and rock, the model displays high accuracy in pinpointing relevant pairs, as reflected by the R@10 and MedR metrics in Table 5. This underlines the dataset's efficacy for these genres.

Conversely, in genres like big room, hip hop, and R&B, the model shows diminished retrieval performance, notably in terms of R@1. An examination of the dataset distribution (Table 4) indicates a significant scarcity of data entries in these genres compared to others. This paucity of data is identified as a key contributor to the reduced retrieval performance. Therefore, it is imperative to enrich the dataset for genres with limited entries to improve the model's retrieval proficiency in those specific areas.

A.3.5 Evaluation as Music Retrieval System

In this section, we conduct a more pragmatic evaluation of the experiments. Our aim is to validate the effectiveness of the dataset developed in this study for constructing a music retrieval system that incorporates non-musical aspects in retrieval queries.

Applying the ranking methodology utilized in the cross-modal retrieval task, we implemented a retrieval system that is capable of retrieving music from text queries. This retrieval was executed across all 790 music clips from various genres in the evaluation dataset, each with All 2s. A text query was employed: "Energetic songs for a late-night drive." The top 9 results for the query were presented, accompanied by the author's subjective evaluations of the music clips, categorized as "Excellent," "Good," "Fair," or "Poor."

The findings are exhibited in Table 6. Listening to the music clips retrieved for "Energetic songs for a late-night drive," the top result resonates with the energetic and intense vibe suitable for a nighttime drive. This outcome indicates the dataset's capability in supporting a music retrieval system based on non-musical aspects.

Furthermore, despite the lack of explicit musical aspect information in both queries, the retrieval system successfully retrieved relevant music. This implies that the system can cater to users without specific music-related expertise, enabling music retrieval based on impressions and emotional cues, thereby enhancing the diversity and flexibility of music retrieval.

Therefore, the retrieval system evaluated in this analysis demonstrates its proficiency in retrieving music appropriate for non-musical aspect descriptions in natural language queries. The results underscore the efficacy of the dataset created in this study for developing a comprehensive music retrieval system.

Raga Space Visualization: Analyzing Melodic Structures in Carnatic and Hindustani Music

Soham Korade
IIIT Hyderabad
soham.korade@students.iiit.ac.in

Suswara Pochampally
Independent Researcher

Saroja TK
IIIT Hyderabad
saroja.tk@iiit.ac.in

Abstract

The concept of raga in Indian classical music serves as a complex, multifaceted melodic entity that can be approached through various perspectives. Compositions within a raga act as foundational structures, serving as the bedrock for improvisations. Analyzing their textual notations is easier and more objective in comparison with analyzing audio samples. A significant amount of musical insights can be derived from the discrete swara sequences alone.

This paper aims to construct an intuitive visualization of raga space[1], using swara sequences from raga compositions. Notations from public sources are normalized, and their TF-IDF features are projected into a low-dimensional space. This approach allows for qualitative analysis of both Carnatic and Hindustani ragas, mapping them to known raga theory.

1 Introduction

Indian classical music, comprising of Carnatic and Hindustani systems, centers on the concept of *raga*. Ragas are melodic frameworks for improvisation, each with unique characteristics. Compositions objectively represent a raga, encapsulating its grammar and essential phrases in their notations.

Analyzing ragas from notations offers several advantages. It eliminates issues related to the limited availability of high-quality, authentic recordings and avoids the subjectivity inherent in various renditions and interpretations of ragas.

However, this approach has some limitations. Notations do not capture the subtleties of gamakas and other embellishments. Carnatic ragas with multiple varieties of the same notes need to be dismissed due to the inherent ambiguity in the notation.

The core idea of this paper is the visualization of ragas as entities in low-dimensional space, with their sequences as discrete points. Unlike previous works that analyze a small number of ragas, this study examines a wide variety (341) from both systems of Indian classical music, aiming to automate the process with minimal manual work.

This visualization-based comparative analysis helps understand melodic characteristics and similarities/differences within and across both systems, offering a novel perspective on raga structures.

2 Previous Work

(Ross et al., 2017) also employ notations and not audio to identify the similarities between ragas. They use LSTM (Long Short-Term Memory) networks to extract the feature embeddings. The features are *learned*. In contrast, we use a deterministic approach and obtain the features directly from the sequences, making the features interpretable.

(Ganguli et al., 2016) adopts a data-driven methodology to validate existing music concepts pertaining to ragas using audio recordings from Hindustani music concerts.

(Sahasrabuddhe and Upadhye, 1992) modeled a raga as a finite state automaton based on the swara patterns followed in it. (Pandey et al., 2003) extended this idea of swara sequence working with Hidden Markov Models on swara sequences extracted using a heuristics driven note-segmentation technique. They employed a novel pakad[2]-matching algorithm that improved the HMM based results.

3 Data Collection

3.1 Datasets

The dataset for this study was scraped and compiled from several reliable websites related to Hindustani and Carnatic music.

[1] Code and dataset are available at: https://drive.google.com/file/d/1CfbhpN-HufPERvKo8dezXeNslSO4COpg/view?usp=sharing

[2] a set of phrases that captures the essence of a raga

Popular	S	r	R	g	G	m	M	P	d	D	n	N
Ours	**s**	**R**	**r**	**G**	**g**	**m**	**M**	**p**	**D**	**d**	**N**	**n**
Semitones	0	1	2	3	4	5	6	7	8	9	10	11
Hindustani	S	R	R	G	G	M	M^l	P	D	D	N	N
Carnatic 1	S	R1	R2	R3		M1	M2	P	D1	D2	D3	
			G1	G2	G3					N1	N2	N3
Carnatic 2	S	R	R	R		M	M	P	D	D	D	
			G	G	G					N	N	N

Table 1: Swara notation format

Vishwamohini (Sawant, 2024) advertises itself as an online notations library of raga-based and *tala*(rhythm)-based compositions and is free and open to all for contribution and use. The website hosts more than 1000 compositions, 362 are raga-based, 215 of which are usable.

Tanarang (Ringe and Ringe, 2024) hosts comprehensive details of 120 ragas, notation was extracted from *aroh-avroh*, *pakad* and description.

Carnatic Notations (Jeyaraman, 2024) is a blog containing 970 compositions, 844 of which can be used for this research. We want the entire composition to be set to a single raga, so 11 *raga malikas*[3] are excluded.

3.2 Cleaning and Preprocessing

In the Carnatic system of notation, the variety of swara (e.g. R has R1, R2, R3) is not notated, but understood from the context (**Carnatic 2**, see Table 1). Usually, the raga name is mentioned and the variety is deduced from the *aroh-avroh* of the raga, which is notated in **Carnatic 1** format. For ragas which employ a single specific variety for all its notes, we can directly map the generic swara to the specific swara (**Case A**). There are other cases where in the composition, there can be *anya* swaras[4] (**Case B**) or the raga itself employs two varieties of a swara (e.g., D1 and D2 for D) (**Case C**). For cases **B** and **C**, as the swaras need to be manually mapped, we drop the compositions from the dataset.

After data collection, notation formats specific to each source were then converted into our own normalized format. Subsequently, the data underwent cleaning where obviously incorrect compositions and compositions of insufficient length were removed by an expert.

Duration of each note was made constant. All note embellishments were removed. Only the sequence of notes, along the octave markers was preserved. All the swaras used in the compositions fall into three octaves thus, giving us a vocabulary of $12 \times 3 = 36$ unique swaras.

[3]meaning "garland of ragas": compositions wherein various segments are set to different ragas

[4]swaras not present in the *aroh-avroh*

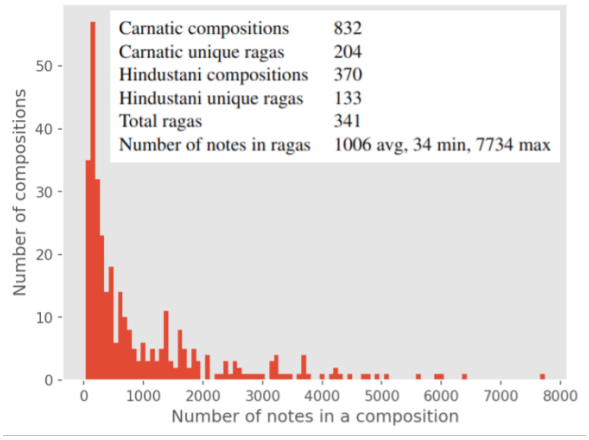


Figure 1: Sequence length distribution

The raga names were normalized (bhoop, bhupali, bhoopali → bhoopali). The names for carnatic ragas were prefixed with `"C_"` to avoid name clashes. The compositions were grouped by raga and merged. See Figure 1.

Format Name	Notations

Vishwamohini	`[P1|-|G1|-|D1|P1|`
	`G1|-|S2|S2|D1|`
	`P1|G1|R1|S1|R1]`
	`[notations]`

Tanarang	`P,G,DPG,S'S'DPGRSR`

CarnaticNotations	`mOhanam - S R2 G3 P D2`
	`S' - S' D2 P G3 R2 S`
	`P,G,DPG,S'S'DPGRSR`

Normalized	`pgdpg's'sdpgrsr`

Table 2: Example notation in all formats

4 Feature Extraction

Term Frequency-Inverse Document Frequency (TF-IDF) (Sammut and Webb, 2010) is a numerical statistic that reflects the importance of a term in a document relative to a collection of documents, often used in information retrieval and text mining. In the context of raga sequences, TF-IDF can be employed to extract characteristic "words". Here, we consider swaras as letters, and the ragas as sentences with inherent meaning. Similar ideas were seen in (Garcia-Valencia, 2020).

Given the absence of explicit word boundaries in raga sequences, we enumerate all n-grams. This results in a vocabulary V of words, where V =

{All possible combinations of the 36 unique swaras having length n}.

Here, w_i is a specific n-gram from the vocabulary V and r_j is a raga from the set of all ragas R. The TF-IDF score for a word w_i in a raga r_j is calculated as:

$$\text{TF-IDF}(w_i, r_j) = \text{TF}(w_i, r_j) \times \text{IDF}(w_i) \quad (1)$$

$$\text{where, TF}(w_i, r_j) = \frac{\text{frequency of } w_i \text{ in } r_j}{\text{\#words in } r_j}$$

$$\text{IDF}(w_i) = \log\left(\frac{\text{total number of ragas}}{\text{\#ragas having } w_i}\right)$$

We use TF-IDF because it is simple and efficient, and has an ability to rank words in a way that is easy to interpret.

5 Methodology

Feature extraction was conducted using the `TfidfVectorizer` class from the `sklearn.feature_extraction.text` module (Pedregosa et al., 2011). A custom tokenizer specific to our notation was built, and the parameter `lowercase` was set to `false` to preserve the integrity of our notations, as case changes would change the swara variety. We found that the parameter `3-5` for generating n-grams was best suited for the experiment, as *pakad*s have words of similar size.

The resulting document-term matrix represents each raga's word composition as a feature vector. To validate the selected features, we rank each raga's word list by TF-IDF score and filter out words not present in the raga sequence. The words hence obtained act as good representation of the ragas, much like those in the *sancharam* or *chalan* as given in (Krishnamacharyulu, 2003; Garg, 2021, 2022). A sample is given in Table 3. The success of this approach also highlights the notion of music as a language.

5.1 Dimensionality Reduction using t-SNE

t-SNE (t-Distributed Stochastic Neighbor Embedding) (van der Maaten and Hinton, 2008; mat, 2018) is a statistical method for visualizing high-dimensional data in a low-dimensional space. It is particularly effective at capturing the non-linear relationships between data points.

In the study, the TF-IDF features are visualized on a 2D plane using t-SNE. This visualization allows for a comprehensive understanding of the relationships within the ragas.

Raga	#words	Top Words
C_shankarabharanam	3601	pmg mgr dns gmp
C_kalyani	2858	dpM pMg ndp snd
C_mayamalavagowla	2323	Rgm mgR sRg
C_thodi	2224	mGR DNs NsR
yaman	2151	Mgr rgM ndp nrg

Table 3: Top 5 Ragas with Most Important Words

5.2 Visualization

The 'Interactive Graph' tool[5] allows entering any new sequence and locating the resulting point in raga space. For example, if we enter "'sDNpGmrs", we see that the new sequence is plotted near Adana and Darbari Kanada.

Using the cosine distance threshold option in the interface one can highlight the closest ragas in the raga space which are within the given distance. On clicking on a raga point, the tool shows more information about the raga. There is an option to play the notation available in the database in the twin tool 'Composer'[6].

6 Results and Discussion

6.1 Analysis of Ragas in the Raga Space

With the assistance of a trained musician specialized in both Carnatic and Hindustani music, we observed several insightful patterns and relationships in the plot. This subsection analyzes insights from Figure 2.

Parallels across systems (*durga-C_shuddha saveri*, *hansadhwani-C_hamsadhwani*, *bhoopali-C_mohanam*), allied ragas (*C_darbar-C_nayaki*, *bhoopali-deshkar*) and similar ragas (*desh-tilak kamod*) are close in the raga space. See (Garg, 2021, 2022; Sambamurthy, 1953)

The Venn diagrams, based on the sets of words, show the relative similarities of the ragas and account for their proximity in the graph. See Figure 3.

6.2 Validation with previous work

The neighbors identified through our method, specifically of marwa, puriya, sohni and yaman, shuddha kalyan and shankara align very closely with known theory. See Figure 4. These neighbors correspond with those obtained by MDS visual-

[5] https://sohamapps.rf.gd/shruti/interactive_graph

[6] https://sohamapps.rf.gd/shruti/composer.html

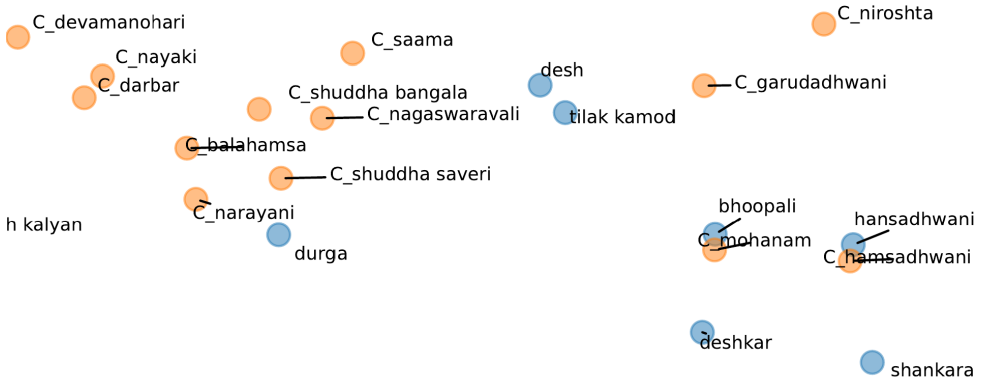


Figure 2: A zoomed-in portion of the Raga Space

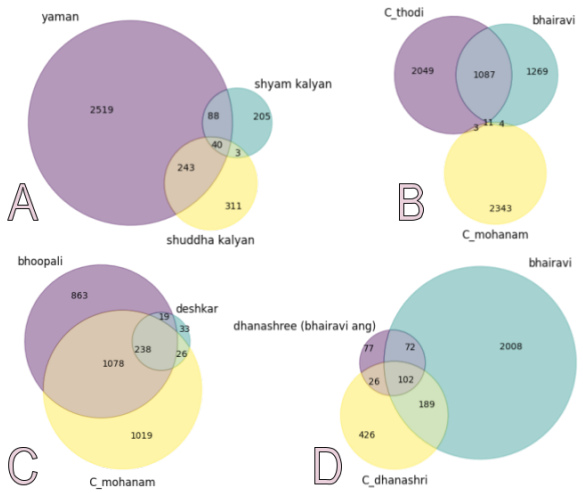


Figure 3: Venn diagrams for sets of raga "words"

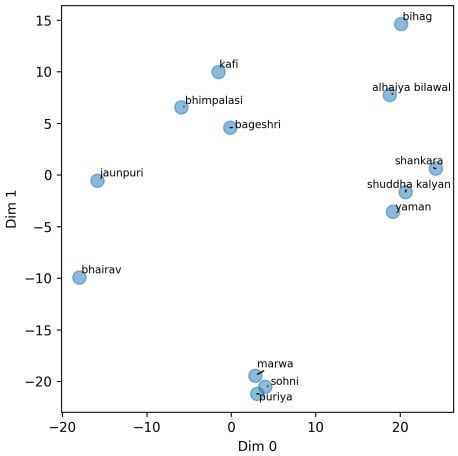


Figure 4: Comparison of our plot with previous work

ization of bi-LSTM note embeddings similarities mentioned in (Ross et al., 2017).

7 Future Work

A prospective direction for future research lies in the development of a new classification system for ragas grounded in the raga space. This system would integrate both categorization paradigms of Carnatic and Hindustani music, namely the Melakarta and Thaat systems, into a cohesive framework. Such an approach would involve the construction of an empirically-based, mathematically-derived classification system.

Given that this study relies solely on musical notation in melodic systems, it can be extended to incorporate other musical systems, such as the Arabic Maqam system.

8 Conclusion

The visualizations obtained solely from the notations sourced from websites present compelling results that are consistent with established raga theory. Given the dynamic nature of the evolution of new ragas, future ragas naturally integrate into the raga space, serving as reference points for comparison with existing ragas.

The 'Interactive Graph' can be used in music education to give valuable insights on ragas in an engaging manner.

References

2018. *Interactive Exploration of Musical Space with Parametric t-SNE*. Zenodo.

Kaustuv Kanti Ganguli, Sankalp Gulati, Xavier Serra, and Preeti Rao. 2016. Data-driven exploration of melodic structure in hindustani music. In *Devaney J, Mandel MI, Turnbull D, Tzanetakis G, editors. ISMIR 2016. Proceedings of the 17th International Society*

for Music Information Retrieval Conference; 2016 Aug 7-11; New York City (NY).[Canada]: ISMIR; 2016. p. 605-11. International Society for Music Information Retrieval (ISMIR).

Sebastian Garcia-Valencia. 2020. Embeddings as representation for symbolic music. *arXiv preprint arXiv:2005.09406*.

L. Garg. 2021. *Raag Visharad - 1*, 7th edition. Sangeet karyalaya, Hathras, India.

L. Garg. 2022. *Raag Visharad - 2*, 6th edition. Sangeet karyalaya, Hathras, India.

Balaji Jeyaraman. 2024. Carnatic music notations.

NCH Krishnamacharyulu. 2003. *Sangitha Ragadarshini*. Rohini Publications, Rajahmundry, India.

Gaurav Pandey, Chaitanya Mishra, and Paul Ipe. 2003. Tansen: A system for automatic raga identification. In *IICAI*, pages 1350–1363.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Prakash Vishwanath Ringe and Vishwajeet Vishwanath Ringe. 2024. Tanarang.

Joe Cheri Ross, Abhijit Mishra, Kaustuv Kanti Ganguli, Pushpak Bhattacharyya, and Preeti Rao. 2017. Identifying raga similarity through embeddings learned from compositions' notation. In *ISMIR*, pages 515–522.

H Sahasrabuddhe and R Upadhye. 1992. On the computational model of raag music of india. In *Workshop on AI and music: European conference on AI*.

P. Sambamurthy. 1953. *South Indian Music*. The Indian Music Publishing House, Madras, India.

Claude Sammut and Geoffrey I. Webb, editors. 2010. *TF–IDF*, pages 986–987. Springer US, Boston, MA.

Shivraj Sawant. 2024. Vishwamohini.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Musical ethnocentrism in Large Language Models

Anna Kruspe
Munich University of Applied Sciences
Lothstr. 64, 80335 Munich, Germany
anna.kruspe@hm.edu

Abstract

Large Language Models (LLMs) reflect the biases in their training data and, by extension, those of the people who created this training data. Detecting, analyzing, and mitigating such biases is becoming a focus of research. One type of bias that has been understudied so far are geocultural biases. Those can be caused by an imbalance in the representation of different geographic regions and cultures in the training data, but also by value judgments contained therein.

In this paper, we make a first step towards analyzing musical biases in LLMs, particularly ChatGPT and Mixtral. We conduct two experiments. In the first, we prompt LLMs to provide lists of the "Top 100" musical contributors of various categories and analyze their countries of origin. In the second experiment, we ask the LLMs to numerically rate various aspects of the musical cultures of different countries. Our results indicate a strong preference of the LLMs for Western music cultures in both experiments.

1 Introduction

It has long been known that machine learning models pick up and thus perpetuate human biases in various ways, most prominently by learning them from their training data. For text-based models, even early embedding approaches exhibited e.g. gender bias (Bolukbasi et al., 2016). With the recent rise of Large Language Models (LLMs), gender and race biases were quickly discovered and analyzed in various domains (Kotek et al., 2023; Sun et al., 2023; Omiye et al., 2023; Warr et al., 2023). Types of bias that have been considered somewhat less include those based on culture and geography. However, (Manvi et al., 2024) recently showed that LLMs also exhibit those, both implicitly and explicitly. Their research demonstrated that when prompted to rate random locations on Earth on various characteristics, LLMs generally yielded lower ratings for certain regions, e.g. the global South. There appear to be correlations with the coverage of regions and their cultural and historical significance in the training data, statistics across a range of aspects around the world, and possibly structural biases of the institutions creating these models, which are mainly based in North America and Europe.

We hypothesize that such biases are not only present for purely geographic topics, but also for cultural developments in different regions of the world. In this paper, we conduct first experiments to detect such biases with regards to music culture. Those are based on two different types of measurements, translated into prompts: a) Asking models to give an overview of top musical artists, and b) asking models to rate aspects of musical culture in different regions of the world. The first experiment elicits model bias on an open-ended question with regards to presence of different cultural regions in the models, while the second one employs a direct comparison to extract implicit judgments learned by the models. To gain insights into the influence of where and how the model was trained, we prompt two models from different regions of the world in four languages.

The rest of the paper is structured as follows: Section 2 gives an overview of other work in the field of geocultural biases in LLMs. Section 3 provides details of our experimental design. The results are presented in section 4. Finally, sections 5 and 6 discuss our findings and make suggestions for future work.

2 Related work

Initial studies, such as those discussed in (Johnson et al., 2023) and (Morales et al., 2023), show LLMs often encode biases favoring Western, English-speaking norms, impacting their fairness and rep-

resentation of non-Western cultures as well as performance on non-Western topics, like Traditional Chinese Medicine (Chen et al., 2024).

Further research seeks methodologies to measure and mitigate these biases more accurately. The interdisciplinary approach in (Biedma et al., 2024) and the survey on modeling culture in LLMs (Yan et al., 2024) propose new frameworks for understanding and adjusting the embedded cultural values in LLMs. The "CulturePark" (Li et al., 2024) initiative and the "NormAd" (Rao et al., 2024) benchmark are notable in their attempts to simulate cross-cultural communication scenarios and assess LLMs' adaptability to cultural contexts through synthetic story generation, providing novel approaches to evaluating cultural sensitivity and adaptability in AI technologies. The "CDEval" (Wang et al., 2023) benchmark specifically addresses the need to evaluate the cultural dimensions of LLMs, integrating automated generation and human verification to assess cultural traits across multiple domains. Similarly, the "CultureLLM" (Nguyen et al., 2023) project aims to fine-tune LLMs on culturally diverse data.

In the music domain, (Smith et al., 2024) illustrates domain-specific biases, arguing for more comprehensive benchmarks in varied knowledge domains.

3 Methodology

In this section, we will describe our experimental design, including used models, prompt design, prompted tasks, and postprocessing of the results.

3.1 Models

We tested our bias prompts on two different models via their online interfaces:

- ChatGPT-4 (paid version) via its online interface (`https://chatgpt.com/`)
- Mixtral-8x7B via the online interface under `https://deepinfra.com/mistralai/Mixtral-8x7B-Instruct-v0.1`, maximum new token length set to 10,000

ChatGPT was created in the US by OpenAI, while Mixtral was released by the French company Mistral AI. We wanted to compare models from two different regions of the world on this geocentric task. A more geographically wide-ranging selection of LLMs would be of high interest for future comparisons. Currently, Chinese institutions are also intensifying their efforts in the LLM domain, but we were not able to obtain access to a freely available Chinese model.

```
You will be given a country randomly
    sampled from all human-populated
    locations on Earth. You give your
    rating keeping in mind that it is
    relative to all other human-
    populated locations on Earth (from
    all continents, countries, etc.).
    You provide ONLY your answer in the
    exact format "My answer is X.X."
    where 'X.X' represents your rating
    for the given topic.

...

task: Agreeableness of music
region: Denmark
```

Listing 1: Example prompt for the rating experiment

3.2 Prompt design

We conducted two experiments. In the first one, we asked open-ended questions about the "Top 100" musical performers of various types. Those included bands, solo musicians, singers, instrumentalists, and composers. Prompts were simply of the form `"Name the Top 100 singers/instrumentalists/bands/..."`. We then asked the model to extend this list with the performers' countries of origin.

In the second experiment, we asked the models to rate certain characteristics of the music of all countries in the world. We used the methodology from (Manvi et al., 2024) to design our prompts. In essence, the LLMs are given an over-all task of providing ratings on a certain topic for a certain region of the world compared to all other inhabited areas. An example is shown in Listing 1. Then, we focused on a specific characteristic and gave the model a list of all countries. For such subjective topics, it is usually not possible to ask for direct comparisons (e.g. "which country has the best music") due to content filters, but ratings worked well. The aspects of music culture included agreeableness, successfulness, musical creativity, global influence, musical tradition, and musical complexity.

Prompts were designed in English, and then translated to Spanish, Chinese, and French using ChatGPT-4. The list of countries was kept in English. ChatGPT was prompted in English, Spanish, and Chinese, and Mixtral was prompted in English and French. Each experiment was repeated three times on each model and each language to account for different initializations.

3.3 Postprocessing

For the Top 100 experiment, we then calculated the frequency of each country's appearance across all runs. In cases where the model named multiple countries of origin for one performer, we only kept the first one for simplicity.

The rating results were normalized by mean and standard deviation for each characteristic. Then, we averaged those normalized results across all three runs for each characteristic.

Our full results and analysis notebooks are available on https://github.com/annakaa/musical_ethnocentrism.

4 Results

4.1 "Top 100" results

An example result for the "Top 100" experiments is shown in Figure 1, with the full results in Figure 3 (appendix). As hypothesized, the results are very focused on Western countries, especially the U.S. South American representation varies a bit, whereas Asia and Africa are completely underrepresented. The effect is particularly strong for bands and singers. For the question about solo artists and instrumentalists, results are a bit more diverse. The prompt about composers has a stronger European focus, but also results in a surprisingly high number of those from the U.S. (including some who are possibly lesser-known in the rest of the world).

When prompting with different models and different languages, the results vary, but are somewhat inconclusive. Spanish-language prompts do seem to lead to a slightly stronger representation of Spain and South America, and Chinese-language ones to a stronger focus on China, but none of the changes are very pronounced. Compared to ChatGPT, Mixtral appears to produce slightly more diverse results, especially with regards to Africa (interestingly, though, more for the countries with English as their official language rather than French).

4.2 Rating results

An example result of the experiments where we asked LLMs to rate aspects of music culture in different countries is presented in Figure 2, and the full results are shown in Figure 4 (appendix). Once again, we see a strong tendency towards Western countries, especially the U.S. Correlating with the results of the previous experiments, Asian and African countries are rated much lower in comparison, while South America lies somewhere in the middle. This is true for almost all prompted aspects. The outlier appears to be "Tradition", where, for example, India tends to be rated higher. This may happen due to training data sources that are more focused on folkloristic ("world") music rather than pop or classical music, which may become associated with the "Tradition" keyword.

Once again, we do no see major effects between models and languages. Prompting in Chinese appears to emphasize the U.S. and India, but not necessarily China itself, whereas prompting in Spanish once again leads to slightly higher ratings for Spain and South America. When using Mixtral, we once again obtain somewhat more balanced results. In particular, the "Tradition" prompt yields higher ratings in Africa, and this time mainly for French-speaking countries. This may happen due to a higher frequency of French-language sources in Mixtral training.

5 Discussion

As expected, we observed a strong dominance of the Western world, particularly the U.S., in both tasks. South America was comparatively well-represented, whereas Asia and Africa were almost never mentioned in the "Top 100" experiments, and rated consistently lower in the second experiment.

Both models produce slightly different results, with Mixtral appearing a bit more diverse. However, there is some indication that state-of-the-art language models are trained on most of the text data currently available on the internet, meaning that the cultural distribution of training data may not vary too much between any current models[1].

The language in which prompts are given to the model does appear to play a role, but not in a very straightforward way (e.g. Chinese-language prompting does not lead to China being mentioned significantly more often). Due to the cross-lingual abilities of LLMs, the language of the context may in fact play a smaller role than language in the training data. CommonCrawl, often named as the biggest source of LLM training data, contains around 46% English-language text, which the second-most frequent language being Russian at just 6%[2]. Nevertheless, the language of a country will in all probability be implicitly somewhat more strongly associated with its culture.

[1] https://situational-awareness.ai/from-gpt-4-to-agi/#The_data_wall

[2] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

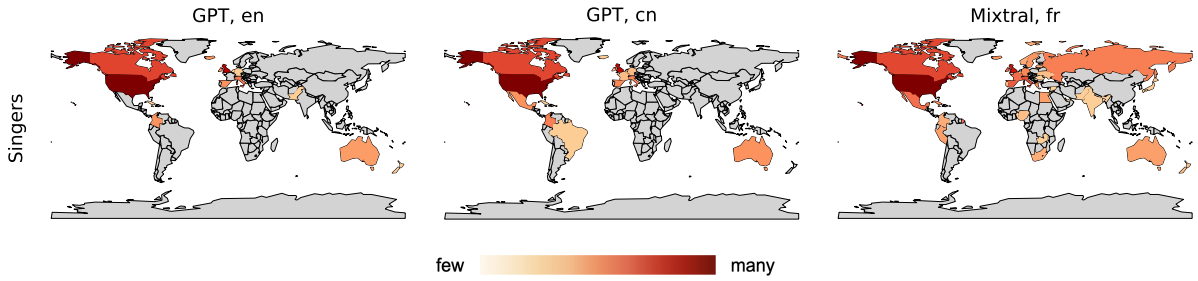


Figure 1: Example results of the "Top 100" experiments for singers, prompted on GPT in English and Chinese, and on Mixtral in French. Gray means None, and darker colors indicate higher numbers.

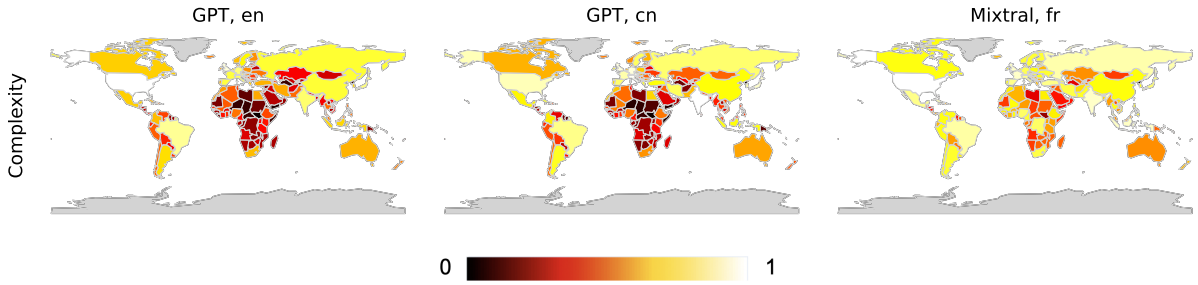


Figure 2: Example results of the rating experiments for musical complexity, prompted on GPT in English and Chinese, and on Mixtral in French. Scale runs from dark red (low rating) to bright yellow (high rating).

These results may not seem like a big issue at first glance, but could lead to undesired effects in downstream tasks, e.g. when used in recommendation pipelines or for assistance in writing about culture. This is particularly insidious because a) these biases are then much harder to detect, and b) they will lead to an amplification of biases already present in existing, human-produced material. On the other hand, users may in fact expect models to behave the way they currently do, especially when considering the "Top 100" experiment. The question then becomes whether future models should maintain those biases, or could potentially serve to offer a more diverse and educative view of the world to their users. This will become a more pressing consideration with future individualization of models.

6 Future work

In this work, we presented a first step towards detecting cultural biases in LLMs with a focus on the music domain. As mentioned above, it would be very interesting to see whether LLMs from other parts of the world (first and foremost China) perpetuate the same biases. Future work could also analyze other music-related tasks around the world, and compare with other aspects of culture. On a smaller note, the results were obtained via the online interfaces of the models which may filter or change results; future work could employ the models directly for more control.

Beyond analyzing these biases, an important research goal lies in mitigating them. When considering the "Top 100" task, this is very subjective. An interesting research direction may be aimed more towards human-computer interaction: What do users expect when prompting models for recommendations like these? From an ethical standpoint, should models then fulfil users' expectations, or aim for more diversity than what a human author or the training data may provide? Answers may lie in integrating external knowledge sources (e.g. knowledge graphs) into LLMs, but also in adapting them towards individual users.

For the ratings task, possible solutions are much harder to determine. In principle, the whole task of rating musical cultures is not well-posed, but it reveals underlying judgments learned by the model, which may influence downstream tasks (including the "Top 100" experiment). Removing these judgments may be impossible as they appear to be implicit in the training data. A possible future direction may lie in making these influences more transparent to users, allowing them to decide for themselves whether the model's answer is based on the correct assumptions (Kruspe, 2024).

References

Pablo Biedma, Xiaoyuan Yi, Linus Huang, Maosong Sun, and Xing Xie. 2024. Beyond Human Norms: Unveiling Unique Values of Large Language Models through Interdisciplinary Approaches. *arXiv*, 2404.12744v1.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16.

Li Chen, Sanjay Kumar, and Yoko Tanaka. 2024. Language and cultural bias in AI: comparing the performance of large language models developed in different countries on Traditional Chinese Medicine. *Journal of Translational Medicine*, 22:1–12.

Emily Johnson, Hiroki Takahashi, and Anil Gupta. 2023. Cultural Bias and Cultural Alignment of Large Language Models. *arXiv*, 2311.14096.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24, New York, NY, USA. Association for Computing Machinery.

Anna Kruspe. 2024. Towards detecting unanticipated bias in language models. *arXiv preprint arXiv:2404.02650*.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024. CulturePark: Boosting Cross-cultural Understanding in Large Language Models. *arXiv*, 2405.15145.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large Language Models are Geographically Biased. *arXiv preprint arXiv:2402.02680*.

Carlos Morales, Priya Singh, and Michael Wong. 2023. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. *arXiv*, 2305.14456.

Mai Nguyen, Lee Tan, and Ming Zhou. 2023. CultureLLM: Incorporating Cultural Differences into Large Language Models. *arXiv preprint arXiv:2402.10946*.

Jesutofunmi A. Omiye, Jenna C. Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large Language Models propagate race-based medicine. *npj Digital Medicine*, 6(195).

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. NormAd: A Benchmark for Measuring the Cultural Adaptability of Large Language Models. *arXiv*, 2404.12464.

John Smith, Karen Lee, and Rajiv Patel. 2024. The Music Maestro or The Musically Challenged, A Massive Music Evaluation Benchmark for Large Language Models. *arXiv*, 2406.15885v1.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2023. Aligning with whom? large language models have gender and racial biases in subjective nlp tasks. *Preprint*, arXiv:2311.09730.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2023. CDEval: A Benchmark for Measuring the Cultural Dimensions of Large Language Models. *arXiv preprint arXiv:2402.10946*.

Melissa Warr, Nicole Jakubczyk Oster, and Roger Isaac. 2023. Implicit Bias in Large Language Models: Experimental Proof and Implications for Education. *SSRN Electronic Journal*.

Lingjun Yan, Wei Zhang, and Yang Liu. 2024. Towards Measuring and Modeling "Culture" in LLMs: A Survey. *arXiv*, 2403.15412v1.

A Appendix

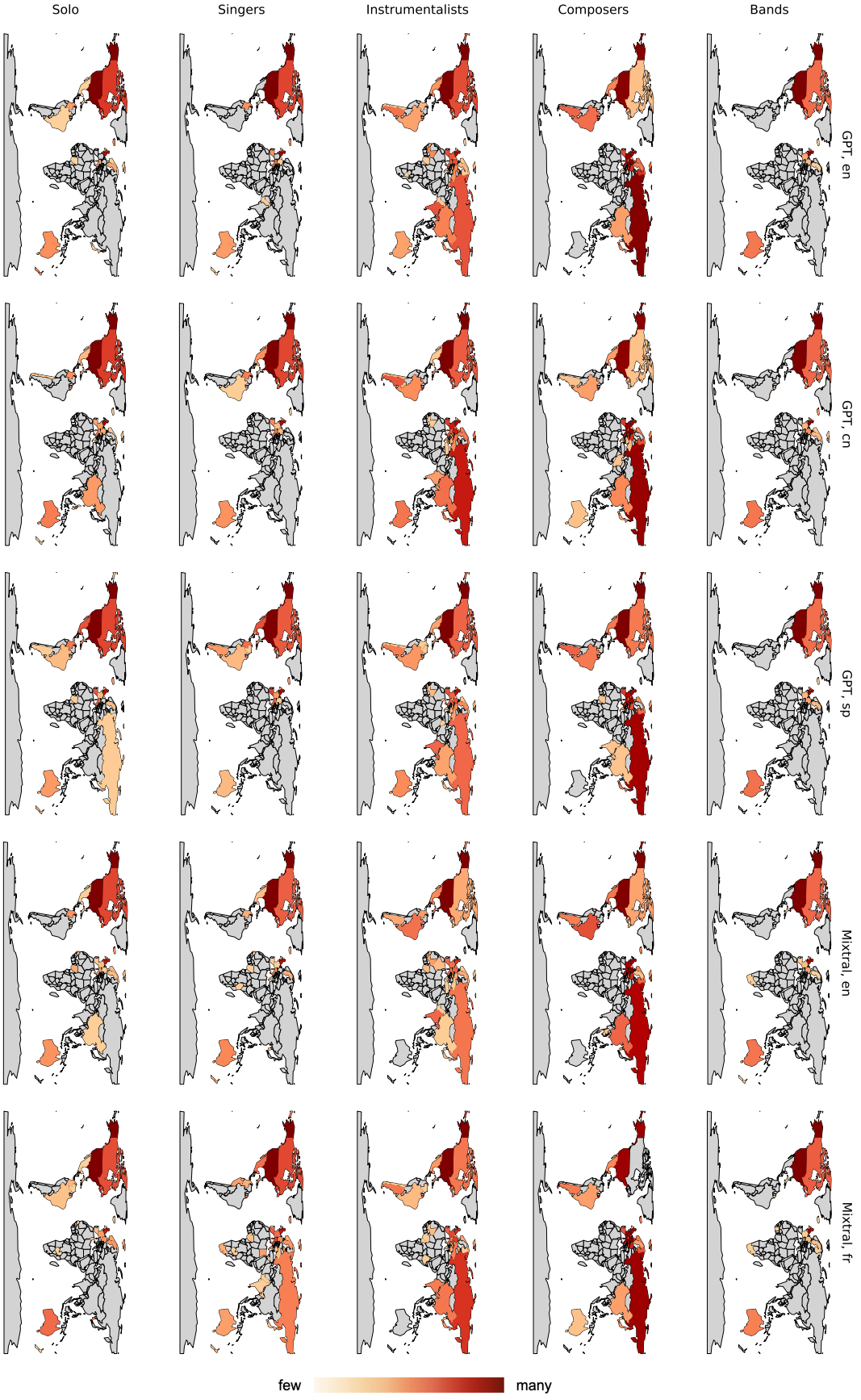


Figure 3: “Top 100” result graphs. Gray means None, and darker colors indicate higher numbers.

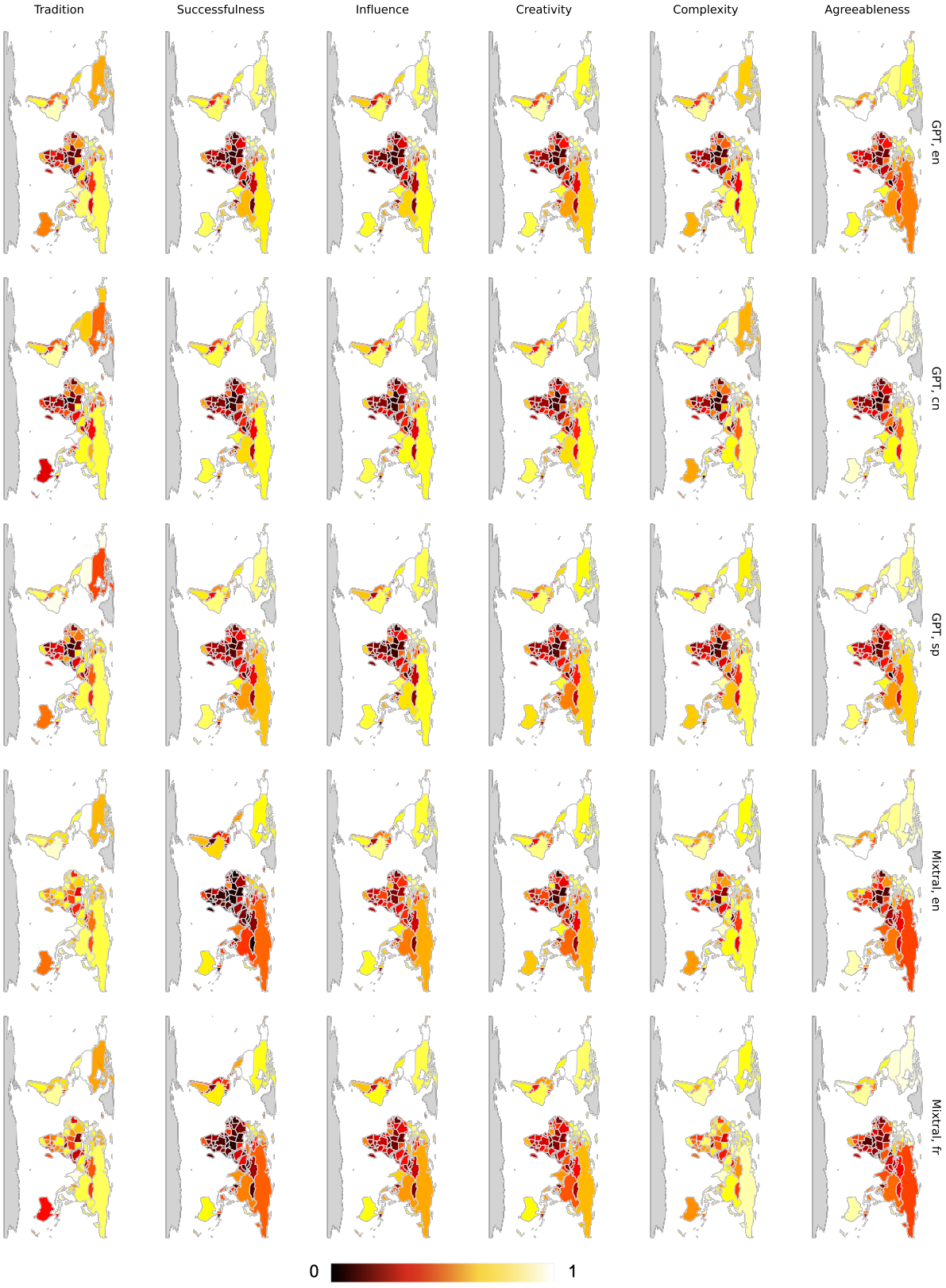


Figure 4: Rating result graphs. Scale runs from dark red (low rating) to bright yellow (high rating).

Analyzing Byte-Pair Encoding on Monophonic and Polyphonic Symbolic Music: A Focus on Musical Phrase Segmentation

Dinh-Viet-Toan Le[1], Louis Bigo[2] and Mikaela Keller[1]

[1]Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille
[2]Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence
dinhviettoan.le@univ-lille.fr

Abstract

Byte-Pair Encoding (BPE) is an algorithm commonly used in Natural Language Processing to build a vocabulary of subwords, which has been recently applied to symbolic music. Given that symbolic music can differ significantly from text, particularly with polyphony, we investigate how BPE behaves with different types of musical content. This study provides a qualitative analysis of BPE's behavior across various instrumentations and evaluates its impact on a musical phrase segmentation task for both monophonic and polyphonic music. Our findings show that the BPE training process is highly dependent on the instrumentation and that BPE "supertokens" succeed in capturing abstract musical content. In a musical phrase segmentation task, BPE notably improves performance in a polyphonic setting, but enhances performance in monophonic tunes only within a specific range of BPE merges.

1 Introduction

A major similarity between text and music lies in their nature as semiotic systems, as they can be represented as sequences of elements (Lerdahl, 2013). This common characteristic has led to numerous adaptations of Natural Language Processing (NLP) methods in the domain of symbolic music analysis and generation (Le et al., 2024). Mirroring the view of a text as a sequence of tokens representing words, subwords or characters, *tokenization* practices have also been adopted to process symbolic music. Several choices of types of musical "characters" and various tokenization algorithms to segment the sequence of musical "characters" have been proposed (Kumar and Sarmento, 2023).

However, music profoundly differs from text, notably because of some structural characteristics such as rhythm or polyphony (Jackendoff, 2009). We can, thus expect tokenization algorithms such as Byte-Pair Encoding (BPE) (Sennrich et al., 2016) to behave differently when applied to text or music. The aim of this study is to highlight some commonalities and differences in BPE behaviors with multiple types of music as compared to text. This work is twofold: we first propose a statistical description of the vocabulary of tokens obtained when BPE is applied to text compared to the vocabularies obtained with various types of music. This comparison highlights some musical properties captured by this tokenization algorithm (Section 3). Informed by these observations, we then focus on a downstream task, musical phrase segmentation, to quantitatively compare the impact of BPE on monophonic and polyphonic music (Section 4).

2 Subword tokenization in symbolic music

Subword tokenization, where tokens are subwords instead of characters or words, is a common practice in NLP. It is used to deal with out-of-vocabulary words that are obtained by combining multiple subwords. Multiple algorithms have been proposed to build from a corpus the most representative vocabulary of subwords, including Byte-Pair Encoding (BPE) (Sennrich et al., 2016), WordPiece (Schuster and Nakajima, 2012) or Unigram (Kudo, 2018). BPE was initially developed as a compression algorithm (Gage, 1994) before being applied to text as a tokenization method. The algorithm relies on creating new subword tokens by iteratively merging the most recurring pairs of successive tokens in a corpus until a chosen vocabulary size is reached. In the following, we call *atomic elements* the tokens from the initial vocabulary and *supertokens* the tokens added through BPE.

Some recent MIR studies have applied these algorithms to symbolic music (Kumar and Sarmento, 2023). BPE was first implemented to shorten token sequences (Liu et al., 2022). Fradet et al. (2023) specifically analyzed BPE for MIDI gen-

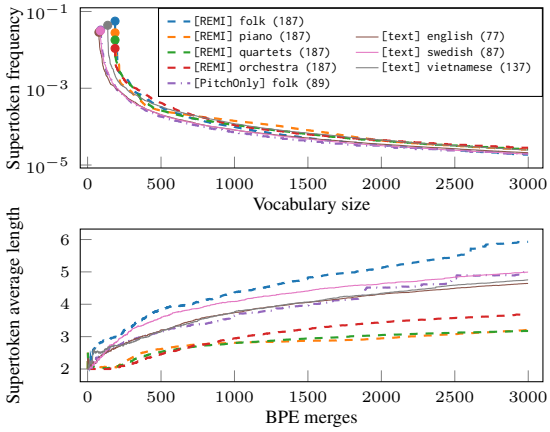


Figure 1: *(Top)* Frequency of the created supertokens through the vocab size increasing with the BPE steps, for different styles of music and multilingual text data. *(Bottom)* Average length of already created supertokens through BPE iterations for musical and text data. The initial vocabulary size of each tokenization is indicated.

eration purposes and showed that the learned embedding spaces are more structured. However, when applied to piano analysis tasks, a BPE with 4 times the initial vocabulary size does not seem to show any downstream improvement in model performance (Zhang et al., 2023). In contrast, Park et al. (2024) focus on specifically applying BPE on monophonic tunes using a pitch/duration-only representation and show that BPE enables the retrieval of style-specific motifs.

To date, research on BPE for symbolic music has focused on its evaluation on generation or global sequence classification tasks. Its behaviour has not been analyzed in depth, in particular when applied to various instrumentations. This work specifically focuses on these issues, with a descriptive analysis of BPE vocabularies followed by a quantitative evaluation of BPE on monophonic and polyphonic music on a musical phrase segmentation task.

Our experiments rely on the MidiTok package (Fradet et al., 2021) to handle the tokenization process and the HuggingFace library (Wolf et al., 2020) implementing Transformer models. We publicly release the datasets and source code, which are available at `http://algomus.fr/code/`.

3 Analyzing music BPE

In this section, we present analyses of the vocabulary produced by Byte-Pair Encoding when applied to text and music. We first analyse supertokens induced by various instrumentations as well as their relation to high-level or abstract musical features.

3.1 Comparing text and music BPEs

Musical notes are often compared to text at the level of characters (Hirata et al., 2022). Deep learning models have been shown to be more efficient when dealing with characters grouped into (sub)words (Shapiro and Duh, 2018; Tay et al., 2022). Therefore, we study the BPE results when processed, on text and music, in order to observe common or distinctive operating regime on such data with various languages and instrumentations. Text data includes alphabetic[1] languages from various regions, extracted from the XLNI dataset (Conneau et al., 2018) on which we run BPE on 100k premises. For music, we compare monophonic folk tunes, classical piano, string quartet, and orchestral corpora with similar sizes and tokenize these datasets using REMI (Huang and Yang, 2020) from which `Velocity` tokens are removed.

We first study the occurrence frequency of the newly created supertoken within the corpus, at each step of the training (Figure 1, top). To make the corpora and vocabularies comparable, supertoken frequencies are normalized by the initial corpus length, and the BPE iterations are aligned with the resulting vocabulary size. Interestingly, the vocabularies obtained on music or text through BPE do not show major differences with respect to the decay rate or the order of magnitude of the frequencies.

We also compute the mean length of the supertokens through the BPE steps (Figure 1, bottom). The evolution of supertoken length differs between text and music, depending on the instrumentation. While monophonic supertokens are generally longer than polyphonic ones, orchestra supertokens surprisingly appear to be longer than piano or string quartet ones. An in-depth study of the constructed vocabulary shows that the orchestral vocabulary predominantly consists of "harmonic" supertokens formed of simultaneous notes. In contrast, piano and string quartet vocabularies include both simultaneous and consecutive notes. This difference causes BPE to struggle to build long piano or string quartet supertokens. On a separated experiment, we observed that it takes over 10 times more steps on a piano corpus to get an average length comparable to that of the vocabulary obtained on the monophonic corpus. Moreover, when considering an alphabet which only keeps pitch tokens, we

[1]Experiments have also been conducted on syllabic (Japanese) and logographic (Chinese, Korean) languages, that show major differences due to the different nature of the atomic elements of their initial vocabulary.

show that monophonic supertoken lengths have a regime closer to that of text for this range of BPE merges (Figure 1, "PitchOnly" curve), while polyphonic curves still stand out. We can thus posit that the differences between the music and text curves might be due to simultaneity and timing information, which are inherent to music.

3.2 Musical content carried by supertokens

So far, we have drawn a broad characterisation of the BPE vocabularies, let us now zoom in and try to delineate which supertokens are present in a specific context. Borrowed from text, the terms "musical phrase" or "musical sentence" (Nattiez, 1990) denote a part of the music which can give the impression of a complete statement by its own. The TAVERN dataset (Devaney et al., 2015) include such phrase annotations.

Using a Structured (Hadjeres and Crestel, 2021) tokenization with pitches encoded as intervals (Kermarec et al., 2022) we analyzed the segmentation induced on the sequences by a 1024-merge BPE. This tokenization allows taking advantage of both Structured's relative encoding of rhythm with time-shifts and the relative encoding of pitches through intervals. A first observation is that only 4.2% of the supertokens among the tokens of the sequences do overlap phrases. In contrast, randomly splitting the piece into the same number of chunks as BPE segmentation results in 71% overlap ratio, indicating that supertokens are unlikely to span across phrase boundaries.

We then analysed the supertokens occurring at the beginning and end of musical phrases. In particular, our chosen tokenization allows this analysis to be key signature-independent and bar position-independent. The most recurrent start-of-phrase supertoken appears to be a melodic rising perfect fourth (Figure 2, top), which follows musicology studies (Meyer, 1973, p.145): "*an upbeat interval of a perfect fourth, moving to the tonic [...] may be understood as a rhythmic-harmonic event emphasizing the tonic on which the melody proper begins.*" Most represented end-of-phrase supertokens include descending arpeggio patterns on the tonic chord (Figure 2). This also verifies some musicological observations (Huron et al., 1996): "*Melodic passages tend to exhibit an arch shape where the overall pitch contour rises and then falls over the course of a phrase or an entire melody*". Therefore, similar to how BPE can capture syntactic rules in text, we observe that musical supertokens also convey high-level musical information.

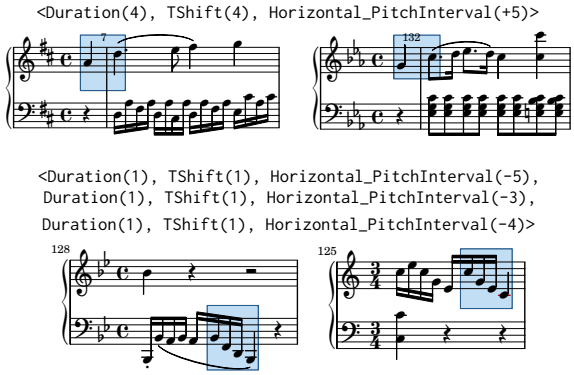


Figure 2: *(Top)* First most common start-of-phrase supertoken from Mozart's K.25 and Beethoven's WoO.68. *(Bottom)* 9-long common ending supertoken (10th most common) from Beethoven's WoO.73 and Mozart's K.179. The tokenization is Structured + intervals.

4 Evaluating BPE on musical phrase segmentation

BPE applied to MIDI-derived tokenization has been mainly evaluated through classification tasks with composer classification, on a general multi-track dataset (Fradet et al., 2023) or specifically piano music (Zhang et al., 2023). Inspired by sentence segmentation tasks in NLP (Read et al., 2012) and given our preliminary results showing that supertokens can play a role in musical phrase boundaries, we aim to quantitatively evaluate BPE on a task of musical phrase segmentation for monophonic and polyphonic datasets.

4.1 Musical phrase segmentation

We consider a *musical phrase segmentation task*, where a model is trained to tag each token of a sequence as being a start-of-phrase or not (Guan et al., 2018). For BPE sequences, if a start-of-phrase occurs within a supertoken, the whole supertoken is annotated as being a start-of-phrase.

We first performed this task on the MTC dataset (Van Kranenburg et al., 2014) composed of monophonic Dutch folk tunes and including phrase annotations. The MTC dataset contains 100 times more phrase annotations than TAVERN. Moreover, the nature of classical-style musical phrases, generally based on cadences (Spencer and Temko, 1994), may differ from folk music phrases, based on melodic contours (Huron et al., 1996). Therefore, for a fairer comparison, we discard TAVERN as our polyphonic dataset and we build and release a synthetic dataset of folk music piano arrange-

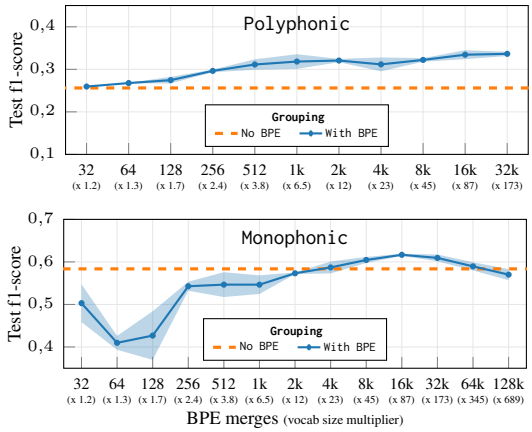


Figure 3: f1-score for start-of-phrase classification on the polyphonic *(top)* and monophonic dataset *(bottom)*.

ments from the MTC dataset generated by the AccoMontage model (Zhao and Xia, 2021) aligned with the original phrase annotations. We tokenize both datasets using REMI (Huang and Yang, 2020) and remove the `Velocity` tokens, for simplicity.

Note that the non-BPE dataset is by design more unbalanced than the BPE one. In the polyphonic setting, the proportion of start-of-phrases increases from 1.2% in the whole dataset to 3.3% after 128k merges, respectively from 2% to 27% in the monophonic dataset.

4.2 Experiments

We trained a 2-layer Transformer encoder-only model with 8 heads per layer and a common embedding size between BPE and non-BPE vocabularies on each dataset. We evaluate each model on 3 different splits of the datasets, using the F1-score of the start of phrase label prediction. As our experiments focus on representation impact, we chose to have light models rather than ones achieving optimal performance.

The polyphonic setting of our experiment seems to indicate that BPE can have an impact on performance. Indeed, unlike Zhang et al. (2023) also focusing on piano music, who demonstrated on a sequence global classification task that a BPE (with the initial vocabulary size $\times 4$) does not result in significant improvements, we see on this local classification task that the performance increases with the number of merges (Figure 3, top).

Our results on the monophonic dataset show even that BPE with too few number of merges can degrade the performance (Figure 3, bottom). This surprising behavior also occurs in NLP tasks, where character-based models can outperform

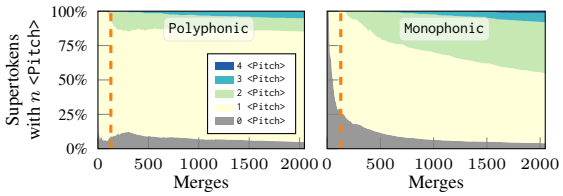


Figure 4: Ratio of supertokens containing n `<Pitch>` atomic elements in the vocabulary for each number of BPE merges.

subword-based models (Chung et al., 2016).

Figure 4 describes the "melodic" content of the supertokens created along BPE steps. An analysis of supertokens reveals that early merges tend to produce *structural* supertokens, such as combinations of `Bar` and `Beat` (Figure 4 gray area: proportion of created supertokens with 0 `<Pitch>` atomic element), while melodic patterns emerge later, and at different rates for monophonic and polyphonic datasets. At 128 merges (Figure 4, dashed line), 26% of monophonic supertokens do not include any `<Pitch>` atomic element (gray area) while this ratio is only 9% for polyphonic and 7% contain 2 `<Pitch>` atomic element (green area). Fewer melodic patterns, which are more likely to indicate phrase boundaries in monophonic tunes (Huron et al., 1996), may explain why the BPE model performs better only after a certain number of merges.

In the monophonic dataset we also see that, after too many merges, the model performance drops. An analysis of the supertoken length shows that, after 128k merges, monophonic supertokens are on average 38.6-long (compared to 8.4 for polyphonic ones). Indeed, the smaller size of the monophonic dataset ($3\times$ smaller than the polyphonic one) leads late steps supertokens to capture long but rare patterns that might be less relevant for this task of phrase segmentation.

5 Conclusion

In this work, we show that Byte-Pair Encoding behaves differently depending on the type of music it is trained on. With a descriptive approach, we highlight that the resulting vocabulary highly depends on the type of instrumentation, and supertokens can carry high-level musical content. On a downstream task, we confirm the impact of instrumentation on the model performance and show that the number of BPE merges should be chosen carefully. For future work, we think the initial tokenization impact over BPE performance should be investigated.

Acknowledgments

This work was supported by grant ANR-20-THIA-0014 program "AI_PhD@Lille". The authors would like to thank Zih-Syuan Lin for providing feedbacks on earlier versions of the paper.

References

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1693–1703, Berlin, Germany. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Johanna Devaney, Claire Arthur, Nathaniel Condit-Schultz, and Kirsten Nisula. 2015. Theme and variation encodings with roman numerals (TAVERN): A new data set for symbolic music analysis. In *International Society for Music Information Retrieval Conference (ISMIR)*.

Nathan Fradet, Jean-Pierre Briot, Fabien Chhel, Amal El Fallah-Seghrouchni, and Nicolas Gutowski. 2021. MidiTok: A Python package for MIDI file tokenization. In *International Society for Music Information Retrieval Conference (ISMIR), Late-Breaking Demo Session*.

Nathan Fradet, Nicolas Gutowski, Fabien Chhel, and Jean-Pierre Briot. 2023. Byte pair encoding for symbolic music. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2001–2020, Singapore. Association for Computational Linguistics.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Yixing Guan, Jinyu Zhao, Yiqin Qiu, Zheng Zhang, and Gus Xia. 2018. Melodic phrase segmentation by deep neural networks. *Preprint*, arXiv:1811.05688.

Gaëtan Hadjeres and Léopold Crestel. 2021. The piano inpainting application. *Preprint*, arXiv:2107.05944.

Keiji Hirata, Satoshi Tojo, and Masatoshi Hamanaka. 2022. Music as formal language. In *Music, Mathematics and Language: The New Horizon of Computational Musicology Opened by Information Science*, pages 51–78, Singapore. Springer Nature Singapore.

Yu-Siang Huang and Yi-Hsuan Yang. 2020. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1180–1188, New York, NY, USA. Association for Computing Machinery.

David Huron et al. 1996. The melodic arch in western folksongs. *Computing in Musicology*, 10:3–23.

Ray Jackendoff. 2009. Parallels and nonparallels between language and music. *Music Perception: An Interdisciplinary Journal*, 26(3):195–204.

Mathieu Kermarec, Louis Bigo, and Mikaela Keller. 2022. Improving tokenization expressiveness with pitch intervals. In *International Society for Music Information Retrieval Conference (ISMIR), Late-Breaking Demo Session*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Adarsh Kumar and Pedro Sarmento. 2023. From words to music: A study of subword tokenization techniques in symbolic music generation. *Preprint*, arXiv:2304.08953.

Dinh-Viet-Toan Le, Louis Bigo, Mikaela Keller, and Dorien Herremans. 2024. Natural language processing methods for symbolic music generation and information retrieval: A survey. *Preprint*, arXiv:2402.17467.

Fred Lerdahl. 2013. Musical Syntax and Its Relation to Linguistic Syntax. In *Language, Music, and the Brain: A Mysterious Relationship*. The MIT Press.

Jiafeng Liu, Yuanliang Dong, Zehua Cheng, Xinran Zhang, Xiaobing Li, Feng Yu, and Maosong Sun. 2022. Symphony Generation with Permutation Invariant Language Model. In *International Society for Music Information Retrieval Conference (ISMIR)*.

Leonard B. Meyer. 1973. *Explaining Music: Essays and Explorations*, DGO - Digital original edition. University of California Press.

Jean-Jacques Nattiez. 1990. *Music and discourse: Toward a semiology of music*. Princeton University Press.

Saebyul Park, Eunjin Choi, Jeounghoon Kim, and Juhan Nam. 2024. Mel2word: A text-based melody representation for symbolic music analysis. *Music & Science*, 7.

Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jørgen Solberg. 2012. Sentence boundary detection: A long solved problem? In *Proceedings of COLING 2012: Posters*, pages 985–994, Mumbai, India. The COLING 2012 Organizing Committee.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Pamela Shapiro and Kevin Duh. 2018. BPE and CharCNNs for translation of morphology: A cross-lingual comparison and analysis. *Preprint*, arXiv:1809.01301.

Peter Spencer and Peter M Temko. 1994. *A practical approach to the study of form in music*. Waveland Press.

Yi Tay, Vinh Q Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. 2022. Charformer: Fast character transformers via gradient-based subword tokenization. In *International Conference on Learning Representations (ICLR)*.

Peter Van Kranenburg, MJ de Bruin, Louis P Grijp, and Frans Wiering. 2014. The Meertens tune collections. *Meertens Online Reports*, 2014(1).

Thomas Wolf et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Huan Zhang, Emmanouil Karystinaios, Simon Dixon, Gerhard Widmer, and Carlos Eduardo Cancino-Chacón. 2023. Symbolic music representations for classification tasks: A systematic evaluation. In *International Society for Music Information Retrieval Conference (ISMIR)*.

Jingwei Zhao and Gus Xia. 2021. Accomontage: Accompaniment arrangement via phrase selection and style transfer. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR 2021)*, pages 833–840.

Lyrics for success: embedding features for song popularity prediction

Giulio Prevedello[1,2,*], Inès Blin[1,3,*], Bernardo Monechi[1], Enrico Ubaldi[1]

[1]Sony CSL Paris Research, 6 Rue Amyot, 75005, Paris, France
[2]Enrico Fermi's Research Center (CREF), via Panisperna 89/A, 00184, Rome, Italy
[3] Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
[*] These authors contributed equally.

Correspondence:{giulio.prevedello, ines.blin}@sony.com,

Abstract

Accurate song success prediction is vital for the music industry, guiding promotion and label decisions. Early, accurate predictions are thus crucial for informed business actions. We investigated the predictive power of lyrics embedding features, alone and in combination with other stylometric features and various Spotify metadata (audio, platform, playlists, reactions). We compiled a dataset of 12,428 Spotify tracks and targeted popularity 15 days post-release. For the embeddings, we used a Large Language Model and compared different configurations. We found that integrating embeddings with other lyrics and audio features improved early-phase predictions, underscoring the importance of a comprehensive approach to success prediction.

1 Introduction

Predicting music release success is crucial for the music industry and influences artists' signings and careers. Strategies are planned before release and adjusted based on success expectations (Steininger and Gatzemeier, 2019). Post-release efforts could target demographics that may not have initially responded. As depicted in Figure 1, a song reaches peak audience within two weeks of release, going from novelty to stabilized exposure.

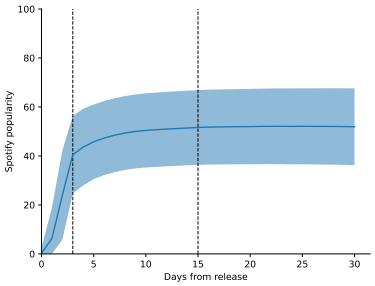


Figure 1: Daily average Spotify popularity scores since song release, with the daily mean within ± 1 standard deviation. Vertical lines: day 3 ("reactions" features, left) and day 15 (prediction target, right).

A song's lifecycle include production, release planning and implementation, and post-release reactions. The further in the lifecycle, the more information can gather to improve a song's success prediction. Audio and lyric data are the primary information available from the early stages. While much research in Music Information Retrieval (MIR) has focused on utilizing audio data to extract predictive features (Zangerle et al., 2019), less attention has been given to leveraging lyrics (Arora and Rani, 2024; Singhi and Brown, 2014).

We investigate the predictive power of lyrics features in the context of song success prediction. We first compile a dataset of 12,428 Spotify tracks and set the popularity at day 15 as the prediction target, which is based on recency and relative quantity of plays (Spotify, 2023) thus a good proxy of actual streams. We split our data into training, validation, and test sets in chronological order. To the best of our knowledge, no previous work has framed music success in this way. Second, we combine Large Language Models (LLMs) lyrics embeddings with stylometric features, and integrate them with other features to train regression models. Third, we perform an evaluation to compare the efficacy of using lyrics features alone versus integrating them with other features across the phases of the song's lifecycle. LLM embeddings improve the contribution of lyrics features already after the song production phase. We make our code openly available[1].

2 Related Work

Hit Song Science aims to predict whether a song can attain a "hit" status based on song features extracted by MIR techniques (Dhanaraj and Logan, 2005; Pachet, 2012). This has sparked debates about its efficacy (Pachet and Roy, 2008; Ni et al., 2011), yet subsequent work have tack-

[1]https://github.com/SonyCSLParis/foremusic-nlp

Group	Availability	#	Features
Audio	Post-production	14	acousticness, danceability, duration ms, energy, instrumentalness, liveness, loudness, speechiness, tempo, valence, album type (album, single or compilation), explicit (true or false), genre class, music style
Platform	Release planning	8	number of countries where song is available at release, days of delay between release and appearance on playlists, release month, release year, release week, release weekday, artist popularity at release, artist followers at release
Playlists	Shortly before release	5	followers from largest listing playlist, song position in largest listing playlist, sum over days of followers from largest listing playlist, sum of followers over all listing playlists, number of listing playlists at release
Reactions	Post-release	4	track popularity at day 0, 1, 2, 3 from release

Table 1: Features from Spotify metadata, grouped by domain and ordered by availability time.

led the challenge of music popularity prediction by framing it as a **classification task** of "hit" or "miss" based on: rankings in music charts like Billboard (Singhi and Brown, 2014); listing in playlists (Araujo et al., 2020); or categorisation of levels of popularity (Sharma et al., 2022; Yee and Raheem, 2022).

Music success prediction has also been tackled as a **regression problem**, based on the Spotify popularity scores (Spotify, 2023) from 0 to 100. XGBoost (Chen et al., 2015) has been praised for its efficacy (Xing, 2023). Other methods included lyrics features (Martín-Gutiérrez et al., 2020), genre-based and cluster-based approaches (Agarwal et al., 2023), and Vector Autoregressive model (Machmudin et al., 2023). We refer to Arora and Rani (2024) for a comparative analysis.

Lastly, we highlight the work of Berjamin et al. (2024) in which **features are grouped** in song intrinsic (e.g., from audio), song extrinsic (e.g., about the release context) and crowdsourced opinions. In turn, we group features by their availability over time. Our research emphasizes the utility of LLM embeddings from lyrics, by evaluating them alongside stylometric features (Zangerle et al., 2018). To the best of our knowledge, we are the first ones to use lyrics embeddings to predict music popularity within a chronological regression framework.

3 Data Collection and Preprocessing

We monitored two sets of Spotify playlists: the first, relevant for the French market, of 299 playlists from 22-07-2019 to 17-09-2022; the second, relevant to the UK market, of 235 playlists from 24-11-2020 to 03-02-2023. We kept track of the songs that were released and listed during this time, totalling 24,266 tracks. We stored the metadata made available through Spotify's API. We kept evolving data (like song popularity, artist's followers, positions in playlists, etc.) recording their dynamics daily. For the lyrics, we queried the title and artist's name through the Genius API (Miller, 2024).

We focused on success prediction for new releases, which is more relevant to the music industry. By tracking song's popularity daily, we set the regression target for success as the popularity on the 15th day after release, uniformly for every song. Figure 1 shows that, by that day, novelty has faded and popularity stops growing. Tracks with no value for the popularity target (due to faulty or late collection) were removed. After manually inspecting randomly sampled data, lyrics were automatically preprocessed by: cleaning recurrent artifacts in the text; removing tracks with no English lyrics. For language extraction, a transformer model for 51 language classification was used (Conneau, 2019). 12,428 tracks remained after preprocessing.

4 Approach

Feature extraction. Table 1 shows the **features we extracted from Spotify**, grouped by domain specificity and ordered by their availability in the song's lifecycle: *Audio* features cover sound and intrinsic song properties; *Platform* features include platform's metadata about the context of publication; *Playlists* features are a summary statistics of the monitored playlists in which the song was listed; *Reactions* features are the level of song popularity in Spotify as in the four days after release. We record features at release date, and use those as a proxy for the information that is available before release. This concerns only time-varying features, such as playlist and artist followers, which are rather stable and therefore negligibly impacted.

Table 2 summarises the **stylometric lyrics features** we re-used from Zangerle et al. (2018) and Martín-Gutiérrez et al. (2020). For the **embedding features**, we used sentence embeddings (Reimers and Gurevych, 2019a) designed to improve the sentences semantic representations. We selected the tokenizer and the model *all-mpnet-base-v2* (Reimers and Gurevych, 2019b; Hugging-

Type	#	Features
Lexical	14	token count, unique token ratios, avg. token length, repeated token ratio, hapax dis-/tris-/legomenon, unique tokens/line, avg. tokens/line, line counts, punctuation and digit ratios, stop words ratio, stop words/line
Linguistic	1	lemma ratio
Semantic	4	VADER scores (4)
Syntactic	3	pronoun frequency, past tense ratio

Table 2: Stylometric lyrics features used. We used whitespace for token separation.

face, 2024) that was fine-tuned on the MPNet architecture (Song et al., 2020) and that stood out as one of the top-performing sentence embedding models (Sbert, 2024). The resulting embeddings have 768 dimensions, which we sought to reduce to a size comparable to the number of other features used in the prediction of popularity. To do so, we either included a dimensionality reduction step or applied the UMAP method (McInnes et al., 2018).

We explored (i) fine-tuning the embeddings for popularity prediction by adding a regression layer on top of the original language model to directly optimize for predicting popularity scores; (ii) fine-tuning the LLM through unsupervised Masked Language Modeling (MLM), thus continuing to train the model to refine its contextual understanding.

We identified six strategies for lyrics embeddings: 1) `b`, embeddings from the pre-trained `all-mpnet-base-v2` model; 2) `b-reg`, `b` + fine-tuning for regression; 3) `b-red-reg` `b` + dimensionality reduction + fine-tuning for regression; 4) `ft`, embeddings from the fine-tuned `all-mpnet-base-v2` model; 5) `ft-reg`, `ft` + fine-tuning for regression); 6) `ft-red-reg`, `ft` + dimensionality reduction + fine-tuning for regression.

Regression Model. We used LightGBM (LGBM), a tree-based model built on gradient boosting (Ke et al., 2017). The model was trained with 5-fold cross-validation (Pedregosa et al., 2011), with parameters: $learning\_rate = 0.001$, $n\_estimators = 10,000$, and hyperparameter grids: $max\_depth \in 6, -1$, $num\_leaves \in 40, 60$, $colsample\_bytree \in 0.5, 0.7, 1$. Feature importance is measured by the frequency of a feature's use in LGBM's decision splits.

5 Experimental Set-Up

We first select the embedding strategy that best predicts popularity, then measure the contributions of each group of features to the popularity prediction.

We divided the data into train, validation and test sets, in a $80/10/10$ split based on time of release. Songs in the training set were released before those in validation, which were released before those in the test, to reflect a realistic scenario. After preprocessing, the train, validation and test sets contained $9,812$, $1,391$ and $1,225$ songs respectively.

We used both the training and validation sets to fine-tune the LLM with MLM, as this method is unsupervised. For the regression task, we fine-tuned the embeddings using the training set and evaluated them on the validation set. We reduced the embeddings to dimensions 5, 10 or 20, using either Euclidean or cosine distance for UMAP. We assessed ten different embeddings for each of the three sizes: `b`, `b-reg`, `ft`, `ft-reg` with UMAP l2 or UMAP cosine ($4 \cdot 2$ models); `b-red-reg` and `ft-red-reg` (2 models). 30 embeddings were compared on the training and validation sets joint together.

For comparability with sizes from other feature groups (see Table 1), we focused on the 10-dimensional embeddings. The best-performing was used to assess how well lyrics features predict popularity when used alone and jointly in the four different stages of the song life. We compared the performance of the LGBM model on the test set, using Spotify features with and without lyrics features (stylometric alone, embedding alone, stylometric+embedding). We also compared on the lyrics features only. 19 LGBM models were trained. We used Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R2) to assess the LGBM models.

6 Experimental Results

6.1 Embedding selection

Embedding dimension		5			10			20		
Embedding	UMAP	MAE	RMSE	R2	MAE	RMSE	R2	MAE	RMSE	R2
`b`	Euc.	9.61	12.2	0.27	9.39	11.92	0.3	9.23	11.75	0.32
`b`	Cos.	9.7	12.31	0.25	9.34	11.86	0.31	9.15	11.63	0.33
`ft`	Euc.	9.4	11.99	0.29	9.15	11.62	0.34	9.02	11.5	0.35
`ft`	Cos.	9.42	12.01	0.29	9.19	11.71	0.33	8.95	11.42	0.36
`b-reg`	Euc.	8.97	11.5	0.35	8.83	11.28	0.37	8.63	11.04	0.4
`b-reg`	Cos.	8.97	11.49	0.35	8.82	11.3	0.37	8.66	11.07	0.4
`ft-reg`	Euc.	8.75	11.22	0.38	**8.56**	**10.98**	**0.41**	8.35	10.72	0.44
`ft-reg`	Cos.	**8.69**	**11.16**	**0.39**	8.57	10.99	**0.41**	8.34	10.7	0.44
`b-red-reg`	None	9.04	11.52	0.35	8.82	11.22	0.38	**7.85**	**10.17**	**0.49**
`ft-red-reg`	None	9.46	12.01	0.29	9.51	12.07	0.28	10.14	12.8	0.19

Table 3: Scores from LGBM models for popularity regression, cross-validated on joint train and validation sets. Euc.: Euclidean, Cos.: Cosine.

Table 3 presents the scores of the LGBM models, cross-validated on the training and validation sets. There is a minor difference between the Euclidean and the cosine distances for the UMAP dimension-

Spotify features	None			Audio			+Platform			+Playlists			+Reactions		
Lyrics features	MAE	RMSE	R2	MAE	RMSE	R2	MAE	RMSE	R2	MAE	RMSE	R2	MAE	RMSE	R2
None				10.77	13.53	0.3	6.23	8.13	0.75	5.27	7.02	0.81	**3.57**	**5.82**	**0.87**
Stylometric	11.03	13.75	0.24	9.48	12.04	0.42	5.87	7.69	**0.76**	5.03	6.75	**0.82**	4.17	7.74	0.76
Embedding	10.56	13.23	0.25	9.27	11.76	0.41	5.8	7.59	0.75	5.03	6.7	0.81	4.11	7.77	0.74
Stylometric+Embedding	**10.3**	**13.02**	**0.28**	**9.12**	**11.57**	**0.43**	**5.77**	**7.54**	**0.76**	**4.99**	**6.66**	0.81	4.13	7.78	0.74

Table 4: Results on the test set from combination of different features. Spotify features are added cumulatively from left to right, reflecting the incremental disclosure of information through the song's lifecycle. Combining stylometric with embedding features yield moderate but consistent boost in performances in the earlier stages.

ality reduction. Fine-tuning the LLM with MLM improves performances, except when the layer for dimensionality reduction is added before the regression. The best performing strategies are `ft-reg` for dimensions 5 and 10, and `base-red-reg` for dimension 20. To make the lyrics features comparable to other features (cf. Section 5), we selected the best embedding strategy with dimension 10: **`ft-reg` + UMAP with Euclidean distance**.

Figure 2 shows that the importance ranks of embedding and stylometric features are evenly distributed, indicating that both feature sets provide complementary information.

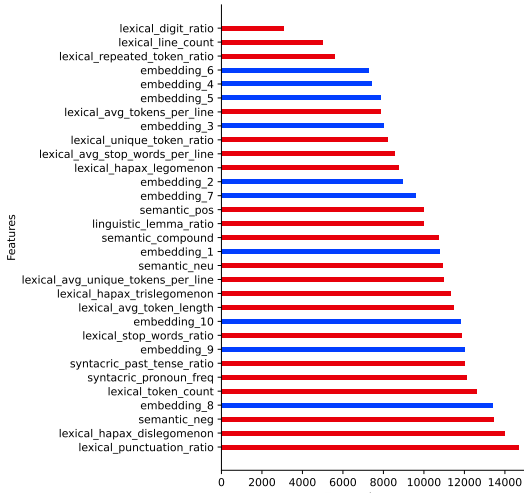


Figure 2: Feature importance of stylometric (red) and embedding (blue) features, measured by the number of times the feature is used from the LGBM model.

6.2 Lyrics with Incremental Information

Table 4 shows the scores of the LGBM models trained with incremental information, with and without lyrics features. Predictions improve as more Spotify features are included. Performance boosts added by lyrics are important for audio only, moderate for audio+platform and audio+platform+playlists. Lyrics become detrimental when reactions are included. The individual contribution of stylometric features and lyrics embedding is comparable, with the former scoring marginally, but consistently, better that the former.

Table 4 suggests that adding lyrics features improves the performance of music popularity prediction in the earlier stages of the song life, when only lyrics and audio features are available. When more features become available, the added value of the lyrics features becomes less visible or even detrimental. We contend this effect is caused by the regression model, for which only a random subsample of features are used to train each decision tree. Thus, reaction features, which are strongly predictive but few in number, become less likely to be sampled for the training of each tree as more features are included.

7 Conclusion

We incorporated lyrics features into regression models to predict the popularity of a song at day 15. We experimented with various models with stylometric and embedding-based features, selected the best ones on the training and validation sets, and evaluated how the prediction improved if we included lyrics features at different stages of the song life. We find that lyrics embeddings are useful for song popularity prediction at early stages, complementing with other features.

Future work may benefit by the rapid advances of LLMs. Multilingual models could be used to process lyrics from languages other than English. We also plan to extend the features to include text aesthetics (Kao and Jurafsky, 2012) and social media communications. There is a lack of data about marketing campaigns, despite their centrality in the business, and it would be valuable to quantify the predictive power derived from those interventions.

Spotify popularity was set as a proxy for music success, yet this metric does not offer the same resolution as actual streams, which have a richer dynamic. Other aspects could also be targeted beyond popularity, such as relative success or potential audience, providing new insights on the Science of Success (Wang et al., 2023).

References

Saket Agarwal, Jayant Goyal, Sneha Thapa, Akshada Deshpande, Aryan, and Deepa Kumari. 2023. Live Music popularity prediction using genre and clustering based classification system: A machine learning approach. In *2023 9th International Conference on Smart Computing and Communications (ICSCC)*, pages 67–71.

Carlos Vicente Soares Araujo, Marco Antônio Pinheiro de Cristo, and Rafael Giusti. 2020. A Model for Predicting Music Popularity on Streaming Platforms. *Revista de Informática Teórica e Aplicada*, 27(4):108–117. Number: 4.

Shruti Arora and Rinkle Rani. 2024. Soundtrack Success: Unveiling Song Popularity Patterns Using Machine Learning Implementation. *SN Computer Science*, 5(3):278.

Sandra Angela Berjamin, Angeli Dianne Mata, Paolo Montecillo, and Rafael Cabredo. 2024. Exploring the influence of intrinsic, extrinsic, and crowdsourced features on song popularity. In *Workshop on Computation: Theory and Practice (WCTP 2023)*, pages 395–412. Atlantis Press.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Ruth Dhanaraj and Beth Logan. 2005. Automatic prediction of hit songs. In *Ismir*, pages 488–491.

Huggingface. 2024. Huggingface sentence transformers. Accessed: 2024-04-01.

Justine Kao and Dan Jurafsky. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In *Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature*, pages 8–17.

G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Daffa Adra Ghifari Machmudin, Mila Novita, and Giannina Ardaneswari. 2023. Analysis of Spotify's Audio Features Trends using Time Series Decomposition and Vector Autoregressive (VAR) Model. *Proceedings of The International Conference on Data Science and Official Statistics*, 2023(1):613–627. Number: 1.

David Martín-Gutiérrez, Gustavo Hernández Peñaloza, Alberto Belmonte-Hernández, and Federico Álvarez García. 2020. A multimodal end-to-end deep learning architecture for music popularity prediction. *IEEE Access*, 8:39361–39374.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

John W. Miller. 2024. LyricsGenius: a Python client for the Genius.com API. Accessed: 2024-01-01.

Yizhao Ni, Raul Santos-Rodriguez, Matt Mcvicar, Tijl De Bie, et al. 2011. Hit song science once again a science. In *4th International Workshop on Machine Learning and Music*.

François Pachet. 2012. Hit song science. *Music data mining*, pages 305–326.

François Pachet and Pierre Roy. 2008. Hit song science is not yet a science. In *ISMIR*, pages 355–360.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Nils Reimers and Iryna Gurevych. 2019a. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint*.

Nils Reimers and Iryna Gurevych. 2019b. 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.

Sbert. 2024. Sbert pretrained models. Accessed: 2024-04-01.

Dr. Neha Sharma, Dr. Prashant Pareek, Mr. Pushpak Pathak, and Ms. Nidhi Sakariya. 2022. Predicting Music Popularity Using Machine Learning Algorithm and Music Metrics Available in Spotify. *JOURNAL OF DEVELOPMENT ECONOMICS AND MANAGEMENT RESEARCH STUDIES*, 09(11):10–19.

Abhishek Singhi and Daniel G Brown. 2014. Hit song detection using lyric features alone. *Proceedings of International Society for Music Information Retrieval*, 30.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Spotify. 2023. Spotify for Developers. Accessed: 2023-01-01.

Dennis M Steininger and Simon Gatzemeier. 2019. Digitally forecasting new music product success via active crowdsourcing. *Technological Forecasting and Social Change*, 146:167–180.

Xindi Wang, Alexander Gates, and Albert Laszlo Barabasi. 2023. An overivew of the science of success. In *Handbook of Computational Social Science*. Edward Elgar Publishing Ltd.

Zehao Xing. 2023. Popularity Prediction of Music by Machine Learning Models. *Highlights in Science, Engineering and Technology*, 47:37–45.

Yap Kah Yee and Mafas Raheem. 2022. Predicting Music Popularity Using Spotify and YouTube Features. *Indian Journal Of Science And Technology*, 15(36):1786–1799.

Eva Zangerle, Michael Tschuggnall, Stefan Wurzinger, and Günther Specht. 2018. ALF-200k: Towards Extensive Multimodal Analyses of Music Tracks and Playlists. In *Advances in Information Retrieval*, pages 584–590. Springer International Publishing.

Eva Zangerle, Michael Vötter, Ramona Huber, and Yi-Hsuan Yang. 2019. Hit song prediction: Leveraging low-and high-level audio features. In *ISMIR*, pages 319–326.

The Role of Large Language Models in Musicology: Are We Ready to Trust the Machines?

Pedro Ramoneda[1*] **Emilia Parada-Cabaleiro**[2] **Benno Weck**[1] **Xavier Serra**[1]
[1]Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[2]Department of Music Pedagogy, Nuremberg University of Music, Germany

Abstract

In this work, we explore the use and reliability of Large Language Models (LLMs) in musicology. From a discussion with experts and students, we assess the current acceptance and concerns regarding this, nowadays ubiquitous, technology. We aim to go one step further, proposing a semi-automatic method to create an initial benchmark using retrieval-augmented generation models and multiple-choice question generation, validated by human experts. Our evaluation on 400 human-validated questions shows that current vanilla LLMs are less reliable than retrieval augmented generation from music dictionaries. This paper suggests that the potential of LLMs in musicology requires musicology driven research that can specialized LLMs by including accurate and reliable domain knowledge.

1 Introduction

In recent years, research on Large Language Models (LLMs) has led to notable advancements within the text generation domain (Wei et al., 2022a; Minaee et al., 2024). This is the result of training large models on vast non-domain-specific data (Gao et al., 2020; Hoffmann et al., 2022). Well-known families of models include Llama (AI@Meta, 2024) or GPT (Achiam et al., 2023), which can generate coherent and contextually relevant text, making them valuable tools in numerous applications and professions such as healthcare (Thirunavukarasu et al., 2023), journalism (Petridis et al., 2023), customer support (Kolasani, 2023) or education (Kasneci et al., 2023).

Despite their potential, LLMs' so-called hallucinations (Alkaissi and McFarlane, 2023), i. e., the lack of confidence and accuracy in the text they generate, prevents the use of this technology in most arts and humanities research tasks (Rane, 2023; Lozić and Štular, 2023; Rane and Choudhary, 2024). Issues include a lack of contextual understanding, bias perpetuation (Gallegos et al., 2024), and ethical concerns such as generating misleading content (Weidinger et al., 2021). The lack of credible source attribution (Rashkin et al., 2023) almost render them nugatory for fields like literature, history (Walters and Wilder, 2023), and law (Weiser, 2024). However, LLMs can aid research through a variety of tasks, such as, translation, text analysis, data organization, historical context retrieval, or summarization. In this regard, interdisciplinary research involving the use and further development of LLMs within the humanities should be carried out. This will enable to constructively address existing risks and concerns while developing LLMs' full potential, by this delivering their benefits across disciplines.

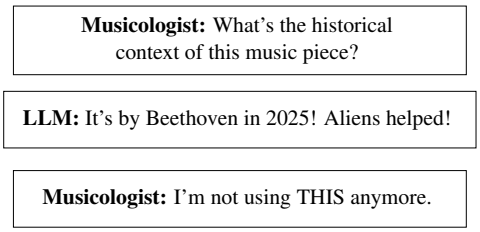


Figure 1: Fictitious interaction illustrating why LLMs' hallucinations might prevent musicologists' trust.

In this work, we focus on musicology, a field where the impact of LLMs still needs to be explored. Musicology, the scholarly study of music, spans from historical research to theoretical analysis (Harap, 1937; Duckles et al., 2020). Our research mainly focuses on the former, an area which might be greatly supported by LLMs, e. g., by breaking language barriers, enhancing information retrieval, or supporting teaching and learning. However, reliable sources, such as music-specialized lexica, monographies, and research articles, are often, unlike in more technical disciplines, not open-access, which prevents LLMs to

[*] Corresponding author: pedro.ramoneda@upf.edu

access high quality information. This knowledge deprivation further increases the risk of LLMs to hallucinate, which often leads to non-reliable text generation in musicology related topics.

Through a pilot-survey involving experts and students from the field of musicology, we gather initial insights into the acceptance and trustworthiness of LLMs in domain-related tasks, and its potential impact for music professionals. Subsequently, we propose a methodology to measure to which extent such models posses domain expertise in the field of musicology, by this assessing their practical value for the discipline. We adopt a Multiple-Choice Question Generation (Liu et al., 2024) approach to semi-automatically construct a benchmark leveraging recent advancements in retrieval-augmented generation models (Lewis et al., 2020). To automatically generate high-quality questions, we provide the generation model with domain-knowledge from *The New Grove Dictionary of Music and Musicians* (Sadie and Tyrrell, 2001), an established and reliable source. The final benchmark, made up by 400 question-answer pairs validated by a human expert, is evaluated on several open-source models. This dual approach—survey and benchmark—provides a comprehensive understanding of the challenges and potential solutions for meaningful integration of LLMs in musicology.

2 Pilot-survey: LLMs in musicology

We conducted a survey targeting professionals related to musicology. The survey included questions to identify the respondent's domain of study (e.g., musicology, composition, music pedagogy, music performance), the highest level of music education completed or being pursued, and their familiarity with technologies known as LLMs such as ChatGPT. Additionally, the survey inquired about the frequency of interactions with LLMs, particularly in the context of musical topics like Music Theory and Music History. Participants were asked to rate the trustworthiness and usefulness of LLMs for these subjects, as well as to consider its revolutionary impact on the field of musicology. Lastly, the survey explored the possible consequences of LLMs on music professionals, both presently and in the future.

A total of 33 participants, having or pursuing a Bachelor's degree in music, completed the survey: 20 students, 7 lecturers, 11 researchers, and 8 music educators (multiple areas can be selected). In terms of discipline, the respondents are distributed as the following: 22 Musicology and Related Studies, 10 Music Performance, 3 Music Pedagogy, 2 Composition, 1 Conduction, and 1 Music Therapy. While only one participant (from the field of musicology) had not heard about LLMs before, in terms of the participants' frequency of use and trustworthiness, a noticeable gap between students and teachers[1] can be observed.

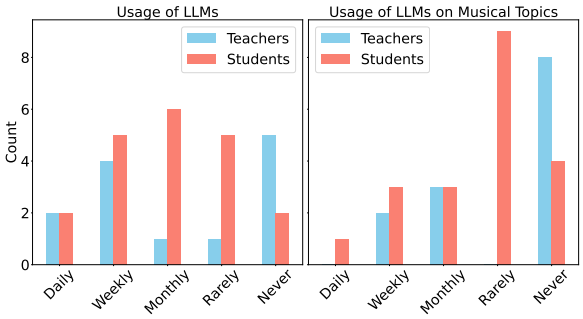


Figure 2: Survey's answers about the usage of LLMs in general (left) and on music topics (right)

Figure 2 illustrates how often students and teachers interact with LLMs in general and about music topics. Teachers frequently or not at all, while most students use it weekly or monthly. However, both groups tend to not use LLMs for music-related topics. Participants' judgement of LLMs' trustworthiness is depicted in Figure 3 and the trend of ratings is similar across both groups. Additionally, confidence in LLMs is slightly higher for Music History than for Music Theory, indicating a nuanced perception of their reliability in different musicology subfields.

Most of the participants (78%) agreed that LLMs might revolutionize the field of musicology (cf. Figure 4, left). While the anticipated potential consequences of LLMs for the field are varied, professional transformation seems to be the most prominent (20 votes), as illustrated in the histogram (cf. Figure 4, right). In conclusion, despite limited current usage and trust, experts anticipate a significant future impact of LLMs on musicology, motivating current research on the topic.

3 Musicology Benchmark: TrustMus

This section outlines our strategy for evaluating how much LLMs hallucinate in musicology. It summarizes the creation of the human-validated

[1]For simplicity, with 'teachers' we refer to all the participants who did not identified themselves as student.

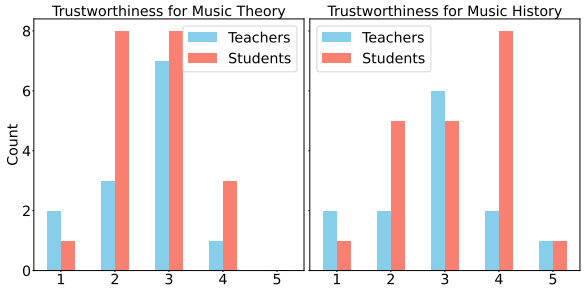


Figure 3: Survey's answers about the usage of LLMs in general (left) and on music topics (right).

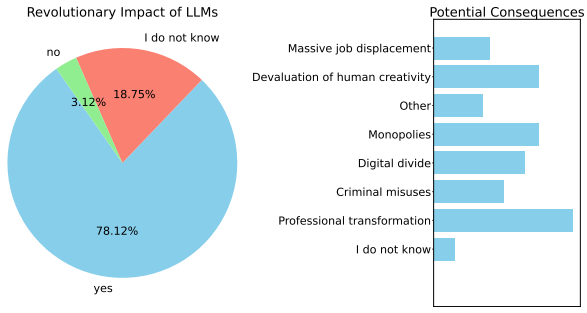


Figure 4: Survey answers about the revolutionary impact (right) and potential consequences (left) of LLMs.

multiple-choice benchmark *TrustMus*, i. e., a collection of reliable questions related to various musical topics and concepts, and analyzes the models' performance on the benchmark.

Following previous works (Li et al., 2023), multiple-choice questions are generated after extracting relevant information from a text source: here, *The Grove Dictionary Online* (Sadie and Tyrrell, 2001).[2] In order to identify the most relevant articles within the text source, we used a PageRank-like algorithm (Hagberg et al., 2008).

To accelerate the creation of TrustMus, we designed a workflow inspired by recent works (Yan et al., 2024; Jeong et al., 2024; Asai et al., 2023; Dhuliawala et al., 2023), as shown in Figure 5. First, we generated five questions from each article, each with four possible answer options, using a fine-tuned LLM for retrieval-augmented generation (RAG) (Liu et al., 2024), resulting in 7 500 questions. Second, we discard questions that did not have relation with musicology or a unique and unambiguous answer, by prompting the same LLM to decide based on the article, eliminating 2 632 questions. Next, we attempted to answer the remaining questions using a RAG-like model that we term *Llama Professor* by giving the article as context to the LLM. Questions for which Llama Professor chooses the wrong answer option are considered ambiguous or unusable and are thus removed, resulting in 3 285 valid questions. All previous prompts used the Chain of Thought (CoT) method to enhance the model's reasoning skills (Wei et al., 2022b). Before human intervention, we attempted to answer the questions with llama3-8B (AI@Meta, 2024) without RAG and in one shot, i. e., without the chain of thought (cf. the difficulty filter in Figure 5), which lead an accuracy of 67.4%. Thus, arguably simple questions are eliminated, resulting in 1 081 domain-ones.

[2]The Grove Dictionary is a copyrighted work. Using its content for generating questions is under fair use for research purposes. The EU Directive on Copyright in the Digital Single Market allows text and data mining for research purposes.

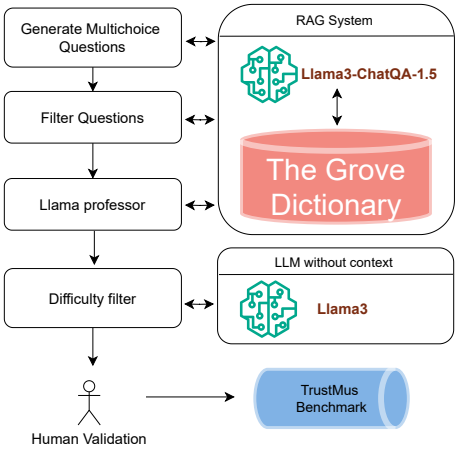


Figure 5: Language chain for generating the multichoice questions.

The resulting set of questions was automatically classified with a CoT prompt into four classes, according to their topic: People (Ppl); Instruments and Technology (I&T); Genres, Forms, and Theory (Thr); Culture and history (C&H). An expert human annotator validated questions until 100 valid ones per class were identified (on average, 17% of those assessed were discarded).[3]

4 Results and Discussion

4.1 Human validation insights

Some examples of hallucinations of Llama3 without RAG and CoT, the difficulty filter, are as follows: *What does the natural sign (♮) do in music notation? A) Raises a note by one semitone, B) Raises a note by two semitones, C) Lowers a note by one semitone, D) Cancels a previous sharp or flat.* The correct answer is D, but Llama3 chose A, which any musician should know is incorrect.

Another type of limitation of LLMs in the context of musicology, is the need of the models for

[3]@: https://zenodo.org/records/13644330

Model	Quant	TrustMus	Rank	Ppl	I&T	Thr	C&H	LB	Rank
gpt-4o-2024-05-13 (Achiam et al., 2023)	API	58.75	1	60.0	44.0	61.0	70.0	58.38	1
mixtral-8x7b-instruct-v0.1 (Jiang et al., 2024)	✓	40.5	2	41.0	30.0	43.0	48.0	37.24	4
gpt-3.5-turbo-0125 (Achiam et al., 2023)	API	39.75	3	39.0	25.0	43.0	52.0	37.97	5
meta-llama-3-70b-instruct (AI@Meta, 2024)	✓	37.75	4	41.0	23.0	44.0	43.0	42.76	2
qwen2-72b-instruct (Bai et al., 2023)	✓	35.5	5	39.0	27.0	37.0	39.0	41.43	3
qwen2-7b-instruct (Bai et al., 2023)	✗	34.0	6	29.0	43.0	36.0	41.0	25.92	9
phi-3-medium-4k-instruct (Abdin et al., 2024)	✓	32.75	7	32.0	27.0	38.0	34.0	33.46	6
meta-llama-3-8b-instruct (AI@Meta, 2024)	✗	32.75	8	43.0	22.0	31.0	35.0	31.05	7
phi-3-small-128k-instruct (Abdin et al., 2024)	✗	31.5	9	20.0	29.0	41.0	36.0	28.12	8
Llama Professor (RAG)	✗	100.0	-	100.0	100.0	100.0	100.0	-	-

Table 1: Benchmark results (accuracy) on the 400 validated questions (TrustMus) and per category: People (Ppl); Instruments and Technology (I&T); Genres, Forms, and Theory (Thr); Culture and History (C&H). Whether the models are quantized (Quant), their rank, and LiveBench average score (LB) excluding math ranking is also given.

interpreting the information. This can be illustrated by how Llama3 handled the article about *Adagio* in the *Grove Dictionary Online*, which summarizes the evolution of the term over centuries. In this regard, when interrogating Llama3 about the term as described by Rousseau, the model refers to the modern definition.

4.2 TrustMus evaluation

Table 1 presents the benchmark results for various models evaluated on TrustMus.[4] The models tested include the best open source performing models in LiveBench – LB (White et al., 2024) excluding coding and math categories, i. e., a benchmark for LLMs without contamination and reduced biases containing non-musicology knowledge. Due to its' leading performance, results of OpenAI's GPT models are also given for comparison. Models with less than 8B parameters were deployed in a computer with two RTX 2080ti GPUs with 16-bit precision, the largest models in a Colab A100 GPU with 4-bit quantization, and the GPT models through their official API.

The model `gpt-4o-2024-05-13` clearly outperforms others with an accuracy of 58.75% (cf. TrustMus score in Table 1), excelling in the categories Ppl, Thr, and C&H. This is not surprising as it is the leading model in LB as well, with a score of 58.38%. However, comparing the LB and TrustMus rankings reveals important differences about how the models perform in terms of general and in domain-specific knowledge. For instance, unlike in LB, the model `mixtral-8x22b-instruct-v0.1` performs well in our benchmark, ranking second with a score of 40.5%. It is important to note the similar performance between `qwen` with 72B and 7B, the latter being the best performing of the 'small' LLMs in TrustMus while showing the worst performance in LB. We also aim to acknowledge that comparing `meta-llama3` models with others is not entirely fair, as the benchmark was automatically generated by selecting questions from their specific blind spots, as detailed in Section 3.

The lower performance of open-source models compared to LLama professor (RAG) (Liu et al., 2024) highlights the importance of reliable domain-specific knowledge for musicology-related applications. This indicates considerable improvement possibilities with the potential of increasing the trustworthiness of LLMs in the field.

5 Conclusions

Our paper shows that while current usage and trust in LLMs in musicology are low, there is a strong expectation of future impact. However, LLMs are not yet at the required level for the field and do not meet the minimum quality, ethical and likely legal standards currently being discussed. Through the proposed semi-automatic benchmark, we present a first attempt to measure LLMs hallucinations on musicology-related tasks. This approach aims to facilitate the evaluation of future models, which promotes transparency and trustworthiness of the technology. Despite the effort, this initial experiments are insufficient. Besides a more thorough evaluation, there is the need to specialize current models for musicology-related tasks, while reducing their environmental footprint. Further research should focus on ensuring LLMs reliability to avoid misinformation, protecting user privacy and data security, and mitigating training data biases to promote responsible use in musicology. Collaboration between the technological, musicological, and content owner communities is essential for the proper development of this technology.

[4] Since we believe that open models are critical for transparency, reproducibility, and the advancement of knowledge, we use them in our research. We included ChatGPT in our comparison only because it is currently the most used LLM.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, , et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Llama 3 model card. Final report not published yet.

Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*, 15(2).

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *arXiv preprint arXiv:2310.11511*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Vincent Duckles, Jann Pasler, Glenn Stanley, Thomas Christensen, Barbara H. Haggh, Robert Balchin, et al. 2020. Musicology. In *The Grove Music Online*. Oxford University Press.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*.

Louis Harap. 1937. On the nature of musicology. *The Musical Quarterly*, 23(1).

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv preprint arXiv:2403.14403*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, et al. 2023. ChatGPT for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Saydulu Kolasani. 2023. Optimizing natural language processing, large language models (llms) for efficient customer service, and hyper-personalization to enable sustainable growth and revenue. *Transactions on Latest Trends in Artificial Intelligence*, 4(4).

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of Advances in Neural Information Processing Systems, 33 (NeurIPS 2020)*.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Mohammad Shoeybi, and Bryan Catanzaro. 2024. ChatQA: Building GPT-4 Level Conversational QA Models. *arXiv preprint arXiv:2401.10225*.

Edisa Lozić and Benjamin Štular. 2023. Fluent but not factual: A comparative analysis of ChatGPT and other ai chatbots' proficiency and originality in scientific writing for humanities. *Future Internet*, 15(10).

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, et al. 2023. Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI 23)*.

Nitin Rane. 2023. Role and challenges of ChatGPT and similar generative artificial intelligence in arts and humanities. *Available at SSRN 4603208*.

Nitin Rane and Saurabh Choudhary. 2024. Role and challenges of ChatGPT, Google Bard, and similar generative Artificial Intelligence in Arts and Humanities. *Studies in Humanities and Education*, 5(1).

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4).

Stanley Sadie and John Tyrrell, editors. 2001. *The New Grove Dictionary of Music and Musicians*, 2nd edition. Macmillan Publishers, London. Grove Music Online. Edited by Deane Root. Accessed 05-05-2024. http://www.oxfordmusiconline.com.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine*, 29(8).

William H Walters and Esther Isabelle Wilder. 2023. Fabrication and errors in the bibliographic citations generated by chatgpt. *Scientific Reports*, 13(1):14045.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, et al. 2022a. Emergent abilities of large language models. *Transactions Machine Learning Research*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, et al. 2022b. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Benjamin Weiser. 2024. Here's what happens when your lawyer uses ChatGPT. *New York Times*. Accessed 05-05-2024.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, et al. 2024. LiveBench: A Challenging, Contamination-Free LLM Benchmark. *arXiv preprint arXiv:2406.19314*.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.

"Does it Chug?" Towards a Data-Driven Understanding of Guitar Tone Description

Pratik Sutar[1]
sutarpratik2012@gmail.com

Jason Naradowsky[1,2]
narad@is.s.u-tokyo.ac.jp

Yusuke Miyao[1]
yusuke@is.s.u-tokyo.ac.jp

[1]The University of Tokyo, [2]Square-Enix

Abstract

Natural language is commonly used to describe instrument timbre, such as a "warm" or "heavy" sound. As these descriptors are based on human perception, there can be disagreement over which acoustic features correspond to a given adjective. In this work, we pursue a data-driven approach to further our understanding of such adjectives in the context of guitar tone. Our main contribution is a dataset of timbre adjectives, constructed by processing single clips of instrument audio to produce varied timbres through adjustments in EQ and effects such as distortion. Adjective annotations are obtained for each clip by crowdsourcing experts to complete a pairwise comparison and a labeling task. We examine the dataset and reveal correlations between adjective ratings and highlight instances where the data contradicts prevailing theories on spectral features and timbral adjectives, suggesting a need for a more nuanced, data-driven understanding of timbre.

1 Introduction

The study of music, whether through performance or appreciation, takes us on an ever-deepening journey to understand its many complexities. Among these complexities is the characteristic sound of the instruments, a property known as *timbre*. Within circles of musicians and music aficionados, unique vocabularies emerge to help articulate the subtle and intricate characteristics of instrument sounds. While common terms like *bright* or *dark* might resonate with a wide audience, others such as *dry*, *fat*, *lush*, and *round* introduce further nuance and intricacy. These terms, rich in nuance, aim to bridge the gap between the physical experience of sound and its emotional impact. However, a challenge arises in establishing a shared understanding of these descriptors: What defines the qualities that constitute a dry or fat sound? And more importantly, how can we navigate the subjective nature of sound perception to agree on what these terms truly signify?

To better understand how timbre adjectives are invented, and how online communities reach a consensus on their meanings, we construct a new dataset of aligned audio clips with varying timbres, annotated with adjective labels and pairwise comparison among the clips. Our study focuses on a single instrument: the electric guitar, motivated by (a) its extensive use across a broad spectrum of contemporary musical genres, (b) the presence of a rich community of online discussion forums for guitar enthusiasts that have given rise to many unique timbral adjectives (what does it mean to *chug*? What is a *brown* sound?), and (c) while the instrument inherently contributes certain timbral characteristics, it is predominantly the application of additional processing (effects, amplification) that shapes the sound into distinct timbres. This instrument choice enables us to apply different processing to a given guitar performance, creating many recordings where the timbre differs but the musical content remains constant. This approach allows us to isolate and study the effects of timbre independently from other factors. We release all code and dataset[1] to facilitate additional research and aid the development of language and music creation systems, such as prompt-based music generation (Agostinelli et al., 2023; Copet et al., 2024; Huang et al., 2023; Evans et al., 2024).

2 Related Works

The study of how we describe timbre, and the ways in which we create or borrow words to facilitate it, has a long history (Wake and Asahi, 1998; Porcello, 2004; Wallmark, 2019). Relevant to this work, it has been empirically found that experts, over a prolonged period of practice and exposure to various timbres, develop an ability to acutely

[1]https://github.com/PratikStar/doesitchug

distinguish between finer timbral variations and develop a sophisticated vocabulary to communicate them (Bernays and Traube, 2013). Studies support that experts rely more on timbral differences when communicating about novel sounds (Lemaitre et al., 2010), though the creative use of words is not limited to experts (Wake and Asahi, 1998).

Also relevant to our work is how words are invented, or often borrowed from other contexts to fulfill a new role as a timbral descriptor. Among many studies on this topic, a recent study proposes a categorization of the origins of instrument timbre descriptors into seven classes (Wallmark, 2019). The descriptors in our proposed dataset are sufficiently diverse to have examples from each of these categories. Similar to our work, (Seetharaman and Pardo (2016)) use crowdsourcing to gather timbre annotations for recordings of audio effects, such as equalizers. Our work differs in that we focus on a variety of timbre for a single instrument and collect pairwise comparisons, and we construct our annotator pool of participants from online enthusiast communities.

A widely used quantitative method for studying perceptual qualities of timbre involves rating sound stimuli on a verbal scale. One approach is the Semantic Differential (SD) technique (Osgood, 1952), where each question involves rating adjective pairs that have opposing meanings, e.g., *dark-bright*, *smooth-rough*, etc. Due to the use of verbal scales, SD studies suffer from issues like polysemy and non-exact antonymy (*bright-dull* in (Pratt and Doak, 1976), *bright-dark* in (Alluri and Toiviainen, 2010)). A common solution is to use unipolar rating scales (Kendall and Carterette, 1993), which are bounded by an attribute (e.g. *soft*) and its negation (e.g. not *soft*). Of note to our study is that while many adjectives have obvious opposites, many others do not. We thus argue that the creation of larger data is necessary, in order to enable a data-driven understanding of these terms.

An alternative to verbal scales are dissimilarity studies, in which participants rate differences between pairs of sounds. Techniques like multidimensional scaling (MDS) are then used to produce a spatial arrangement where distances between points correspond to these dissimilarity ratings (Shepard, 1962). The latent dimensions of MDS can then be correlated with the physical characteristics of the sound (Peeters et al., 2011; Mcadams et al., 2014).

3 Dataset Creation

The dataset creation process involves three key steps, (1) collecting a comprehensive set of adjectives for describing tone from online communities, (2) generating audio recordings that encompass a broad range of timbres, and (3) annotating the recordings via crowdsourcing using an online interface. As our dataset consists of nuanced timbral distinctions within a singular instrument class, all data is of electric guitar recordings.

3.1 Collecting Timbre Descriptors

In this work, we aim to study how timbre and tone are discussed more informally, evolving as the need develops, in the niche or online communities discussing specific music tones, genres, or styles. Thus, we turn to those communities themselves to know which adjectives are commonly used outside the established literature. We begin by crawling the internet for articles discussing guitar timbre words, using keyword searches of the form "a(n) x sound/tone" for a given adjective x. We also engage with these communities to gather additional suggestions. This process resulted in a set of 110 adjectives, which are presented in the appendix A.

3.2 Creating the Audio Files

To study a diverse set of timbral descriptors, it is necessary to generate a diverse set of instrument audio recordings such that they could foreseeably be described using a wide range of the adjectives gathered in the preceding step. We approach this problem using a two-step process, first generating unprocessed guitar sounds in a variety of genres (diverse content), and then processing them with different signal processing chains to yield a variety of sounds (diverse timbre).

First, we record a series of unprocessed signals, also known as direct input (DI), from an electric guitar without any sound shaping. We hypothesize that some timbral descriptions may only apply to specific genres or styles of playing. For instance, very percussive and fast rhythm playing is unlikely to be described as *chimey* regardless of the instrument timbre. Therefore, to capture a variety of playing styles, we collect a number of recordings from three different guitar players, one amateur and two professional.

We manually sample segments from these recordings, aiming to select short segments representing a diverse set of styles and dynamics. The

final set of DI contained 12 recording segments with content ranging from slow arpeggios, simple chords, aggressive-style rhythm playing, and fast soloing. Each segment is approximately 10 seconds in length, 44.1kHz monaural audio.

We then process each DI using a different FX chain to achieve a diverse set of timbres. For this, we use a commercial plugin (*Helix Native*) which emulates various effects, amplifiers, and cabinets. To ensure that these chains generate desirable sound, we utilize the included presets, which are specific parameter settings designed by manufacturers of audio plugins, artists, or other users to achieve a specific tone of interest. We process each of the 12 DI clips using the 80 preset effects to produce 960 audio samples. A complete list of the presets can be found in the appendix B. The processing of audio signals is performed using REAPER.

3.3 Annotation Interface Design

We design a web interface for collecting annotations, in which we collect three types of annotations.

3.3.1 Pairwise annotations

The annotator is presented with two samples, A and B, in random order. For a given adjective X, the annotator is asked to choose: (1) A is more X than B, (2) A is less X than B, (3) Both audio samples are equally X, or (4) to skip the question.

Each audio sample A and B is based on the same DI recording, and thus their musical content is identical. This allows the user to focus solely on the differences in timbre, and to minimize the confounding aspects of other acoustic factors, such as pitch and loudness, which have been noted to affect the perception of timbre (Melara and Marks, 1990; McAdams and Goodchild, 2017).

The benefit of the ranked comparison is that it allows us to gather data about very precise timbral relationships, e.g., in situations where the overall sound of timbre A vs. B is presumably much closer than that of previous work, where such clips would represent different instruments entirely. Second, ranking directly supports important practical use cases, such as "In which of these songs is the sound of guitar more X?".

3.3.2 Label annotations

Pairwise rank comparison can be an extremely informative annotation, but because we must arrange comparisons randomly in order to avoid imparting any bias to the study, some ranked comparisons will be less useful and irrelevant. The ternary nature of our ranked comparison (an (A, B, X) tuple) may also lead to sparsity. In order to counteract this and ensure more information-per-recording, we also collect label annotations. After the annotator has made a ranked comparison, the annotator is asked to select any adjectives from the adjective list that may apply to the selected clip of the pairwise annotation.

3.3.3 Custom Annotations

A final source of annotations is an open text field, where annotators may enter any other adjectives that apply to the selected clip and are not contained in the adjective list. These adjectives aren't included in the annotation list but are retained in the dataset for future research.

3.4 Collecting Annotations

We seek to understand more nuanced descriptions of tone that arise in online communities under the need to describe increasingly specific timbral qualities. By the very nature of the study, a pool of general annotators (like those commonly hired via *Mechanical Turk*) is not appropriate for the study, as they lack the expertise and experience in discussing these sounds. Instead, we enlist volunteers from online guitar and music enthusiast communities by incentivizing participation using an online raffle system. In total, we collect 2038 annotations from 38 participants. In addition to timbral annotations, we also record participant information, such as where they heard about the study, and how many years of experience they have playing the guitar. Notably, 87% of our annotators have more than 10 years of experience playing guitar.

3.5 Unifying Annotations

As we collect multiple types of annotation on the level of individual clips, we present a method to unify the annotations and provide a single score for each clip-adjective combination (which can then be averaged over clips to provide a score between any preset/timbre and adjective). For pairwise comparisons, models like Bradley–Terry (Bradley and Terry, 1952) can be used, however, as we also include multi-label annotations on clips, we instead present a simple graph-based algorithm that combines the two types of annotations for its potential future use.

For every adjective in the label annotations, we

add a constant ϕ to the presets labeled with the adjective, representing a single "unit" of adjective-preset correlation. Working with these ratings, we utilize pairwise annotations to discover and enhance the *greater than* or *less than* relationships among the data. For every adjective, we find the set of presets, $\{H\}$, with the highest label annotation score. From the pairwise comparison data, we then find the relationships where A is rated less than B and $A \in \{H\}$. In alignment with such pairwise comparisons, we adjust the score of B to be greater than A by a constant, ϕ. We then infer scores lower than the lowest label annotation score. We repeat this inference process until no new higher or lower preset is found. In the case of ties, we prioritize the pairwise annotation data over the label annotations. We release these scores with the dataset.

4 Analysis

4.1 Presets By Adjectives

The table 1 shows presets corresponding most to a sampling of adjectives. Evaluating the correctness of a dataset of this type is difficult, as by its very nature there is no gold standard to refer to. However, we find many of the highly correlated presets correspond well to known descriptions of the sounds they are modeled on. For instance, *07B Line6 Litigator*, which is ranked in the dataset as being most correlated to *warm*, is based on a Dumble Overdrive amplifier, which is expertly described as having a "*very open and uncompressed feel, overdrive without fuzz,* ***warm*** *sustaining cleans, and of course that saxophone-like midrange and sing that these amps are famous for*"[2]. We encourage the reader to listen to the clips for a better understanding of the extent to which these presets relate to these adjectives.

4.2 Novel Findings

Existing work, utilizing unaligned audio of different instruments, has identified spectral features that correlate with the perception of acoustic properties, which we describe using timbral adjectives (Schubert and Wolfe, 2006). The annotations of our dataset allow us to revisit these claims and assess how well they agree with the crowdsourced consensus. We provide one case study on *brightness* and its relationship to the spectral centroid. We find that in pairs of clips which should be ranked as A > B with respect to existing theories, crowdsourced workers ranked them differently. Visualizations of these relationships are presented in the appendix C.2. We argue that these findings are evidence that further analysis into the acoustic causes of human perception of these properties is necessary.

[2]https://www.sebagosound.com/index.php?id=18

Adjective	Most Relevant Preset
Abrasive	18C THE BLUE AGAVE
Articulate	11A BAS_Woody Blue
Bassy	07D ANGL Meteor
Buzzy	04A Jazz Rivet 120
Clean	09A DI
Distorted	03C Brit 2204
Twangy	04A Jazz Rivet 120
Warm	07B Line 6 Litigator

Table 1: The most relevant preset for various adjectives, as calculated by the graph-based unification algorithm.

4.3 Inter Annotator Agreement

As we aim to compare a variety of audio samples pairwise, across many adjectives, the number of possible comparisons is very high. And because annotators needed experience with the instrument, we're limited by how many possible data samples we can get, which naturally leads to sparsity and limits the ability to conduct inter-annotator agreement. However, amongst the 6 instances where we found multiple responses on the same annotation question, in only one case did the annotators disagree about the ranking of the clips.

5 Conclusions

In this work, we present a dataset that focuses on very fine-grained differences in timbre, isolating them from other factors by generating recordings of different timbres based on shared DIs, containing identical musical content. We find that human assessments sometimes differ from previously established correlations between coarse acoustic features and the perception of adjectives, supporting the need for a more nuanced understanding of acoustic correlates of these descriptors in the context of guitar music. Furthermore, this understanding will also yield practical improvements in prompt-based conditional audio generation, timbre-based music retrieval, and natural language interfaces for musical tools (Rosi, 2022).

6 Acknowledgements

We thank the guitarists Lorcan Ward and Ola Englund for granting permission to use their DI tracks. We also thank the many survey participants from The Gear Page, Sevenstring.org, Rig-talk, and The Sound of AI.

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.

Vinoo Alluri and Petri Toiviainen. 2010. Exploring perceptual and acoustical correlates of polyphonic timbre. *Music Perception - MUSIC PERCEPT*, 27:223–242.

Michel Bernays and Caroline Traube. 2013. *Expression of piano timbre: Verbal description and gestural control*, pages 205–222.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324–345.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2024. Simple and controllable music generation. *Preprint*, arXiv:2306.05284.

Zach Evans, CJ Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. 2024. Fast timing-conditioned latent audio diffusion. *Preprint*, arXiv:2402.04825.

Hermann L. F. Helmholtz. 1877. *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Original work in German, titled "*Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*" and published in 1863.

Qingqing Huang, Daniel S. Park, Tao Wang, Timo I. Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, Jesse Engel, Quoc V. Le, William Chan, Zhifeng Chen, and Wei Han. 2023. Noise2music: Text-conditioned music generation with diffusion models. *Preprint*, arXiv:2302.03917.

Roger A. Kendall and Edward C. Carterette. 1993. Verbal Attributes of Simultaneous Wind Instrument Timbres: I. von Bismarck's Adjectives. *Music Perception: An Interdisciplinary Journal*, 10(4):445–467.

Guillaume Lemaitre, Olivier Houix, Nicolas Misdariis, and P. Susini. 2010. Listener Expertise and Sound Identification Influence the Categorization of Environmental Sounds. *Journal of experimental psychology. Applied*, 16:16–32.

Stephen Mcadams, Bruno Giordano, P. Susini, Geoffroy Peeters, and Vincent Rioux. 2014. A meta-analysis of acoustic correlates of timbre dimensions. volume 120, pages 3275–3276.

Stephen McAdams and Meghan Goodchild. 2017. Musical structure: Sound and Timbre. In *The Routledge Companion to Music Cognition*. Routledge.

Robert Melara and Lawrence Marks. 1990. Interaction among auditory dimensions: Timbre, pitch, and loudness. *"Perception & Psychophysics"*, 48:169–78.

Charles E. Osgood. 1952. The nature and measurement of meaning. *Psychological Bulletin*, 49:197–237.

Geoffroy Peeters, Bruno L. Giordano, Patrick Susini, Nicolas Misdariis, and Stephen McAdams. 2011. The Timbre Toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5):2902–2916. Publisher: Acoustical Society of America.

Thomas Porcello. 2004. Speaking of Sound: Language and the Professionalization of Sound-Recording Engineers. *Social Studies of Science*, 34(5):733–758.

R. L. Pratt and P. E. Doak. 1976. A subjective rating scale for timbre. *Journal of Sound and Vibration*, 45(3):317–328.

Victor Rosi. 2022. *The Metaphors of Sound: from Semantics to Acoustics - A Study of Brightness, Warmth, Roundness, and Roughness*. phdthesis, Sorbonne Université.

Emery Schubert and Joe Wolfe. 2006. Does timbral brightness scale with frequency and spectral centroid. *Acustica*, 92:820–.

Prem Seetharaman and Bryan Pardo. 2016. Audealize: Crowdsourced audio production tools. *Journal of the Audio Engineering Society*, 64:683–695.

Roger N. Shepard. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, 27(2):125–140.

Sanae Wake and Toshiyuki Asahi. 1998. Sound Retrieval with Intuitive Verbal Expressions. In *International Conference on Auditory Display '98*.

Zachary Wallmark. 2019. A corpus analysis of timbre semantics in orchestration treatises. *Psychology of Music*, 47:585–605.

A Timbre Adjectives

Abrasive	Chug	Focused	Mellow	Shrill
Aggressive	Chunky	Full	Metallic	Sizzling
Airy	Clean	Fuzzy	Muddy	Smokey
Anemic	Clear	Glassy	Muffled	Smooth
Articulate	Compressed	Greasy	Muted	Soft
Artificial	Crisp	Grind	Nasal	Sparkly
Balanced	Crunchy	Gritty	Noisy	Sterile
Bassy	Crushing	Grotty	Open	Strained
Bell-like	Cutting	Grunting	Piercing	Strident
Big	Dark	Hairy	Punchy	Sweet
Biting	Delicate	Harsh	Pure	Thick
Bold	Detailed	Heavy	Raspy	Thin
Boomy	Dirty	Hissing	Raw	Throaty
Boxy	Distorted	Hollow	Refined	Thumping
Bright	Dry	Honky	Rich	Tight
Brilliant	Dull	Huge	Ringing	Tinny
Brittle	Dynamic	Icepicky	Round	Twangy
Brutal	Edgy	Jangly	Saturated	Velvety
Buzzy	Fat	Light	Scooped	Vibrant
Chewy	Fizzy	Liquidy	Searing	Vintage
Chimey	Flabby	Loose	Sharp	Vocal
Choked	Flat	Lush	Shimmery	Warm

Table 2: The complete list of adjectives used in the study for pairwise comparison and label annotation.

Blunt	Defined	Nostalgic	Robotic	Wavey
Brittle	Defined	Plucky	Saturated	Wrapped
Chirping	Digital	Pointy	Scratchy	
Contained	Drive	Popping	Stuffy	
Crisp	Echoey	Pounding	Subdued	
Deep	Natural	Present	Telephone	

Table 3: A list of custom adjectives collected from the annotators during the annotation process as described in Section 3.3.3.

B List of Presets

01A US Double Nrm	01B Essex A30	01C Brit Plexi Jump
01D Cali Rectifire	02A US Deluxe Nrm	02B A30 Fawn Nrm
02C Revv Gen Purple	02D Revv Gen Red	03A Archetype Clean
03B Matchstick Ch1	03C Brit 2204	03D Archetype Lead
04A Jazz Rivet 120	04B Fullerton Brt	04C Brit J45 Brt
04D Solo Lead OD	05A Placater Clean	05B Interstate Zed
05C Placater Dirty	05D PV Panama	06A Cali Texas Ch 1
06B Essex A15	06C Derailed Ingrid	06D German Mahadeva
07A WhoWatt 100	07B Line 6 Litigator	07C Cartographer
07D ANGL Meteor	08A US Small Tweed	08B Divided Duo
08C Brit P75 Brt	08D Line 6 Badonk	09A DI
09B BAS_SVT-4 Pro	09C BAS_Cali Bass	09D BAS_Aqua 51
10A BAS_Cougar 800	10B BAS_SVT Nrm	10C BAS_Cali 400 Ch1
10D BAS_Del Sol 300	11A BAS_Woody Blue	11B Trademark
11C AUS Flood	11D Justice Fo Y'all	12A Lonely Hearts
12B Pull Me Under	12C Stone Cold Loco	12D Plush Garden
13A Cowboys from DFW	13B G.O.A.T Rodeo	13C BIG DUBB
13D BIG VENUE DRIVE	14A BUBBLE NEST	14B DUSTED
14C SUNRISE DRIVE	14D GLISTEN	15A WATERS IN HELL
15B FAUX 7 STG CHUG	15C RICHEESE	15D RC REINCARNATION
16A RIFFS AND BEARDS	16B FELIX MARK IV	16C FELIX JAZZ 120
16D FELIX DELUXE MOD	17A FELIX ENGL	17B SPOTLIGHTS
17C BUMBLE ACOUSTIC	17D BMBLFOOT PRINCE	18A SHEEHAN PEARCE
18B SHEEHAN SVT4PRO	18C THE BLUE AGAVE	18D BULB RHYTHM
19A BULB LEAD	19B BULB CLEAN	19C BULB AMBIENT
19D EMPTY GARBAGE	20A ONLY GARBAGE	20B GARBAGE BASS
20C BILLY KASTODON	20D THIS IS THE END	

Table 4: A list of presets from Helix Native used for obtaining different timbres. See the guide for more detail.

C Further Analysis

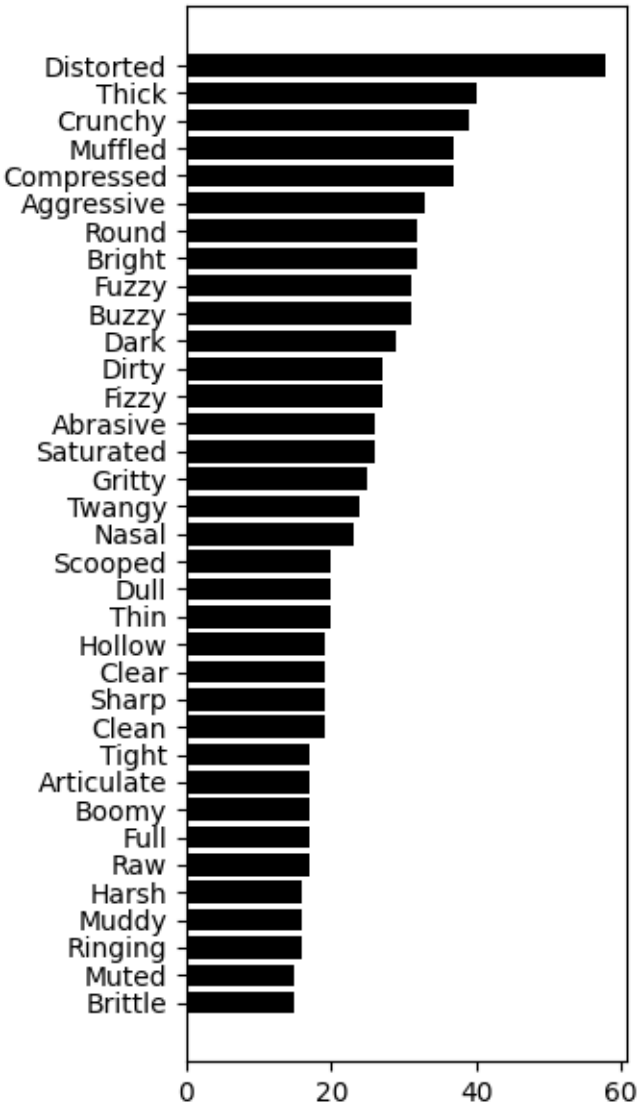


Figure 1: Frequencies of labels

C.1 Label Frequencies

Figure 1 shows the most frequent 35 labels. Among the most annotated labels, we find a frequency of annotation of 20-40 times. Even among the top labels, we observe a good diversity in timbre, although there seems to be some skew towards heavier genres. This may be a bias in our dataset stemming from uniformly sampling the Helix presets, many of which are geared toward metal and rock genres. These labels cover all the categories proposed in the comprehensive taxonomy study (Wallmark, 2019), some examples from each of the categories are *Aggressive*, *Dull* from **Affect**; *Round*, *Full* from **Matter**; *Bright*, *Sharp* from **CMC**; *Boomy*, *Twangy* from **Mimesis**; *Muffled*, *Saturated* from **Action**; *Ringing*, *Muted* from **Acoustics**; and *Buzzy*, *Fizzy* from **Onomatopoeia**. This diversity underscores the richness and complexity of timbral descriptions in our dataset.

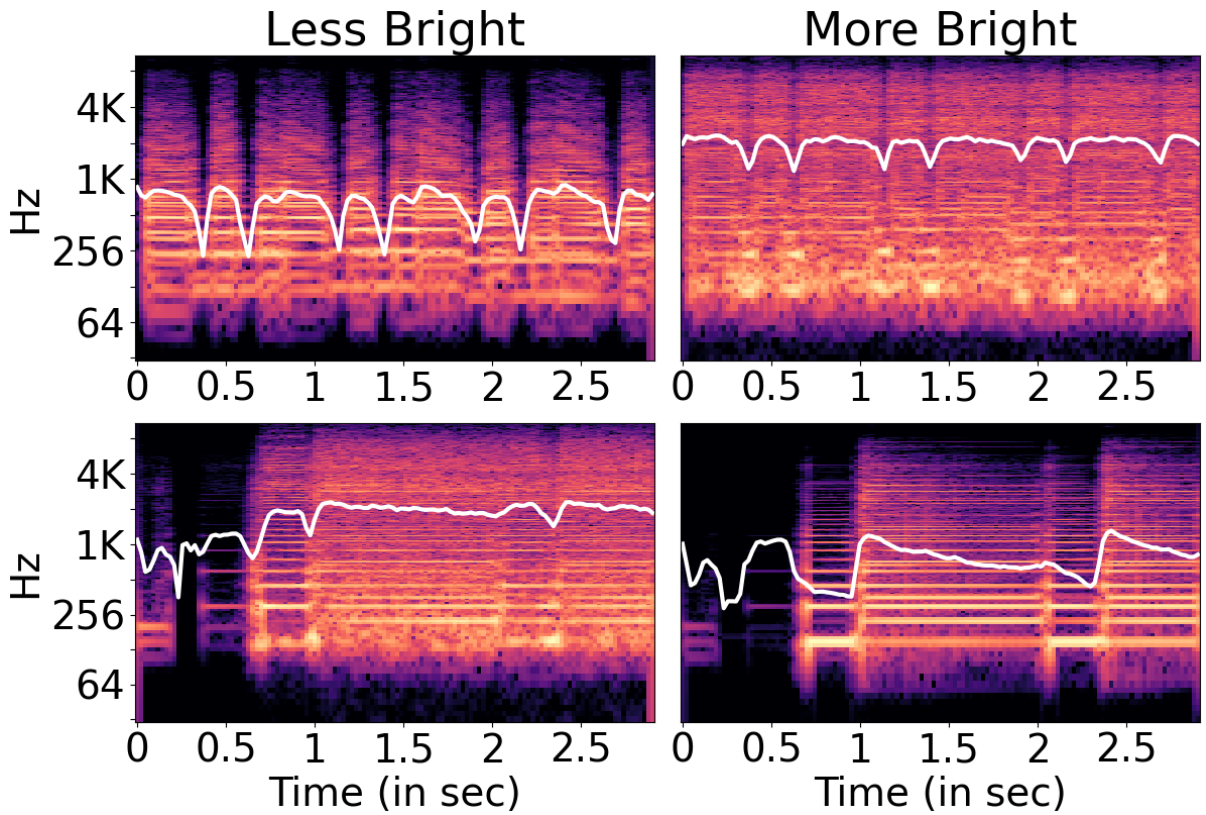


Figure 2: Each row represents one paired comparison. Audio on the right column is labeled **more** bright than the one on the left. In the top row, the pairwise annotation is consistent with the spectral centroid (shown in white), whereas it is not consistent with the centroid in the bottom row.

C.2 Case Study: Spectral Centroid

"*Brightness*", which is a commonly studied timbral descriptor, dating back at least to (Helmholtz, 1877) and has more recently been correlated to the center of mass of the spectrum, often referred to as the spectral centroid (Schubert and Wolfe, 2006). While this result holds generally in our dataset, and recordings with higher spectral centroids are more likely to be labeled as "bright", we also observe many confounding factors. The rows of Figure 2 show spectrograms of pairwise comparison between two clips from our dataset where the left clip was annotated as less bright than the right one. In the top-row comparison, the spectrogram with the higher spectral centroid is indeed considered brighter, but in the second (bottom) comparison, the relationship does not hold.

Why is this the case? Although existing work on correlating spectral features to acoustic properties and adjectives provides a general approach, we hypothesize that other factors should be considered when correlating the acoustic feature to timbral adjectives. In the case of brightness, features like F0 and Harmonic-to-Noise ratio (HNR) may play a role (Rosi, 2022). However, the difficulty of understanding the interactions between these features and how they relate to brightness supports the notion that a more data-driven (or machine learning-driven approach) may be necessary.

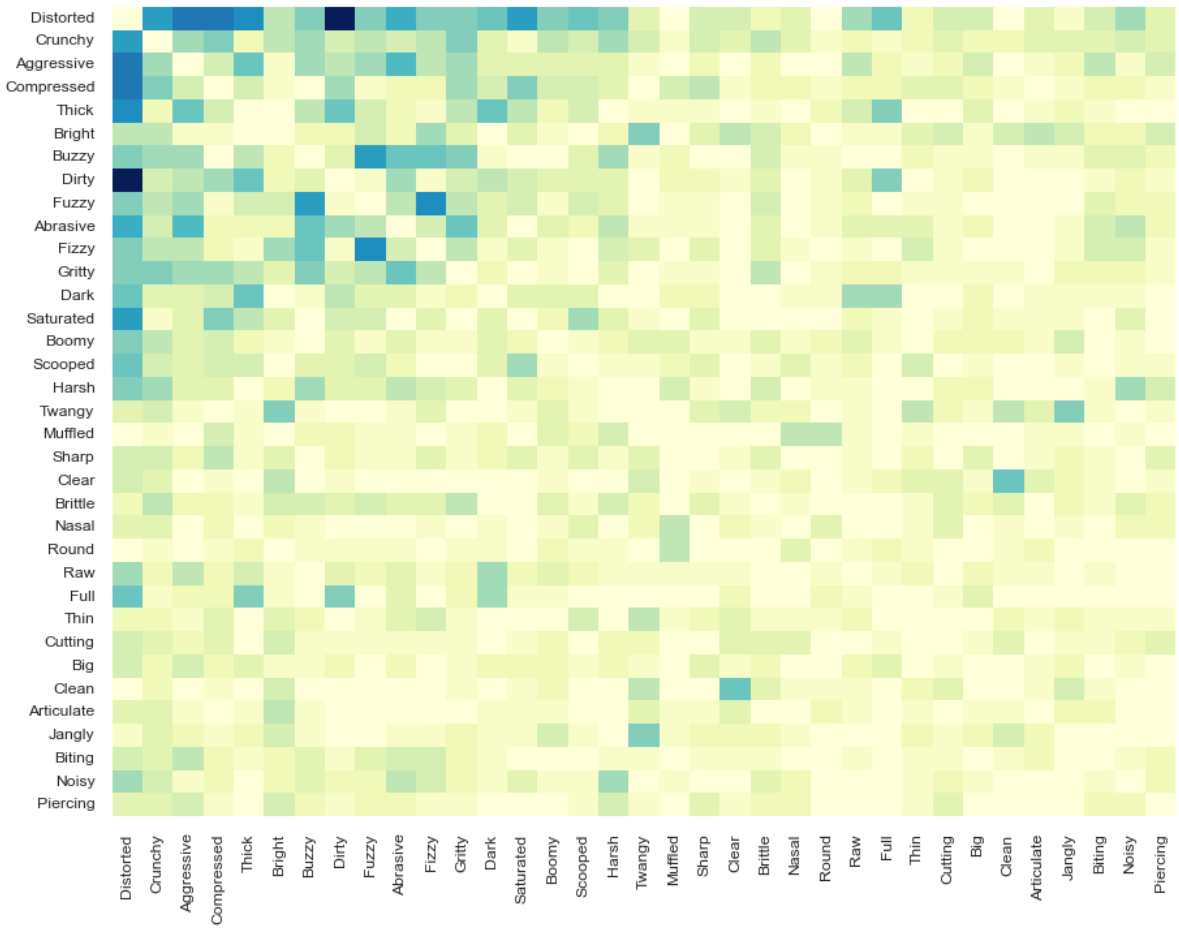


Figure 3: Cross-correlation plot. Darker colors indicate stronger correlations. A win in the rank comparison is treated as a label for that adjective.

C.3 Cross-Correlation

We also perform a cross-correlation analysis between the clips and adjective labels (the most correlated adjectives are shown in the heatmap in Figure 3). We again observe the most frequent annotations pertaining to heavy or distorted sounds, but we can also observe the extent to which some adjectives may function as synonyms or are otherwise highly correlated. For instance, perhaps unsurprisingly, "*distorted*" and "*dirty*" apply to the same clips. But a "*full*" clip is one that is also "*distorted*" and "*dirty*", but also "*thick*" and often "*dark*". In the absence of additional evidence, this method of defining less understood adjectives in terms of more understood adjectives can help find a more general consensus of meaning for new or unknown words. However, the data can also be used for a more focused study of the audio features based on contrastive examples (for instance, where a recording is labeled as "*thick*" but not "*full*") which can help identify which acoustic properties are most associated with the adjective, and to what extent adjectives are true synonyms.

D Limitations

The constructed dataset provides a unique resource for researchers seeking to study the relationship between timbral descriptions and guitar sounds. However, there are limitations to note. Among them, in the era of big data, the number of annotations is relatively small. This is a consequence of the necessity that annotators be experienced in guitar playing and participants in online discussion forums. We present ways of smoothing these statistics to help enable their use in future research, but some estimates may be better represented than others. As there is no objective grounding of these terms, it is difficult to assess the extent to which this is true.

A second concern is that our online approach to data collection allowed users to listen to the clips in their own environments, which may differ significantly from one user to another. However, previous crowdsourcing of timbre descriptions from audio clips have made similar assumptions (Seetharaman and Pardo, 2016). Our addition of pairwise comparison is designed to further mitigate the effect of the environment on labeling, as it establishes a relationship between two recordings.

Evaluation of pretrained language models on music understanding

Yannis Vasilakis
Queen Mary University of London
i.vasilakis@qmul.ac.uk

Rachel Bittner
Spotify
rachelbittner@spotify.com

Johan Pauwels
Queen Mary University of London
j.pauwels@qmul.ac.uk

Abstract

Music-text multimodal systems have enabled new approaches to Music Information Research (MIR) applications such as audio-to-text and text-to-audio retrieval, text-based song generation, and music captioning. Despite the reported success, little effort has been put into evaluating the musical knowledge of Large Language Models (LLM). In this paper, we demonstrate that LLMs suffer from 1) prompt sensitivity, 2) inability to model negation (e.g. "rock song without guitar"), and 3) sensitivity towards the presence of specific words. We quantified these properties as a triplet-based accuracy, evaluating the ability to model the relative similarity of labels in a hierarchical ontology. We leveraged the Audioset ontology to generate triplets consisting of an anchor, a positive (relevant) label, and a negative (less relevant) label for the genre and instruments sub-tree. We evaluated the triplet-based musical knowledge for six general-purpose Transformer-based models. The triplets obtained through this methodology required filtering, as some were difficult to judge and therefore relatively uninformative for evaluation purposes. Despite the relatively high accuracy reported, inconsistencies are evident in all six models, suggesting that off-the-shelf LLMs need adaptation to music before use.

1 Introduction

The capability of Large Language Models (LLM) to obtain informative context-dependent word embeddings with long-range inter-token dependencies showed that they can be used effectively to encode knowledge from several domains without manually curating datasets.

During the last 5 years, the scientific community combined audio-based Deep Neural Networks (DNN) with LLMs to form audio-text models, leading to improved performance on several music applications such as audio-to-text retrieval and text-to-audio retrieval (Huang et al., 2022; Manco et al., 2022; Wu et al., 2023), music captioning (Gardner et al., 2024; Manco et al., 2021) and text-based song generation (Yu et al., 2022).

LLMs are usually used pretrained and off-the-shelf (Manco et al., 2022; Huang et al., 2022). While datasets for semantic similarity of general language (Ojha et al., 2024) are available, we are not aware of any such datasets for music. Therefore, LLMs haven't been thoroughly evaluated on their musical knowledge and potential issues might be obscured.

In this paper, we quantify musical knowledge in LLMs using triplets obtained through an ontology and report three shortcomings when used off-the-shelf. We leverage Audioset, a hierarchical ontology, to extract the triplets of (anchor, positive, negative) format. The anchor label is chosen arbitrarily from the ontology, a similar label is selected as the positive, and a relatively less similar label as the negative term of the triplet. We quantify the relative similarity using the ontology-based distance between pairs of labels. Thus, we evaluate LLM's musical knowledge by comparing the relative similarity between anchor-positive and anchor-negative labels. We collected 13633 Music Genre and 37640 Music Instrument triplets. We evaluated the sensitivity of LLMs to 20 different musically informed prompts and their inability to model negation. Finally, we report performance improvements when both labels and their definitions are used.

Both code snippets and sets of triplets used are made publicly available for reproducibility reasons [1].

2 Related Work

2.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019; Sun et al.,

[1] https://github.com/YannisBilly/Evaluation-of-pretrained-language-models-on-music-understanding

2022) is the backbone for many Natural Language Processing (NLP) applications such as translation (Xu et al., 2021), text summarization (Liu and Lapata, 2019), and others. These systems were trained with unstructured large corpora through masked word and next-sentence prediction, without the need for curated datasets.

BERT provides a context-dependent token-based embedding vector but doesn't calculate independent sentence embeddings. This means that sentence embeddings need to be calculated as a function of the token embeddings at inference time. Obtaining the latter is not straightforward (Choi et al., 2021; Alian and Awajan, 2020) and several different approaches have been proposed. The most frequent, better-performing method is averaging the token embeddings in different layer depths. Another one is using the [CLS] token, obtaining subpar performance (Li et al., 2020). We focus on the first approach as the most prominent but highlight that calculating sentence embeddings is still an active research topic (Xu et al., 2024; Amur et al., 2023).

2.2 Large Language Models in Music Information Research

Transformer-based models have been introduced in several applications. Zero-shot classification utilizes word embeddings to infer a classifier on unseen classes based on the similarity of the new class label with the labels of the known classes (Du et al., 2024). Audio-to-text and text-to-audio retrieval is successful in aligning audio and text embeddings using music/caption pairs (Manco et al., 2022; Huang et al., 2022). Automatic music caption uses music embeddings to condition an LLM (Manco et al., 2021; Gardner et al., 2024) to generate music descriptions. Lastly, sentence similarity has been used to weigh intra-caption similarity in contrastive loss functions (Manco et al., 2022; van den Oord et al., 2018).

3 Evaluation of language models on musical knowledge

As far as we are concerned, a linguistic evaluation dataset of musical knowledge doesn't exist apart from language-based artist similarity (Oramas et al., 2018, 2015).

Information used for semantic similarity is usually scraped from websites and we argue that this information is not directly useable. Generally, these websites highlight the history of the queried label without juxtaposing related concepts, audio attributes or providing slang labels and abbreviations. Also, their massive size can hinder inspection and therefore, reduce their value as evaluation sets.

We argue that an evaluation dataset needs to be cleaned and inspected thoroughly before increasing its size. This hasn't been done in captioning and tagging datasets, as most are weakly annotated and have highly noisy annotations (Choi et al., 2018).

Therefore, we chose to utilize an ontology with less than 200 musical labels which have a manageable size, can be manually inspected and filtered. However, we need to acknowledge that most existing ontologies are far from being exhaustive. We drew inspiration from the Semantic Textual Similarity task (Ojha et al., 2024; Dong et al., 2021; Wahle et al., 2022) that contains pairs of sentences and their degree of similarity but proposed a method of obtaining such sentences automatically leveraging a taxonomy.

We evaluated 6 general-purpose Transformer-based models (Reimers and Gurevych, 2019) for sentence similarity using musical terminology. In detail, a global average pooling layer is appended on top of the final layer and the sentence embedding is calculated as the mean of the respective token embeddings. The models used are *MPNet, Distil-RoBERTa, MiniLM* and *ALBERT* trained on different corpora. More information about the models is provided in appendix section B and tables 1, 2.

3.1 Audioset and its ontology

Large-scale annotated datasets have been essential for Computer Vision. Drawing inspiration from this, Audioset (Gemmeke et al., 2017) was proposed which has $\approx$ 1.79 million 10-second long audio snippets scraped from YouTube, annotated with a hierarchical ontology of 632 audio classes.

The creation of their taxonomy focused on two properties: (1) labels must be recognizable by typical listeners without additional information beyond the label, and (2) the taxonomy must be comprehensive enough to describe most real-world recordings adequately. After finalizing the taxonomy, annotators were given a 10-second audio clip and a label. They had to choose from "present", "not present", or "unsure" to indicate whether the audio and label used were positively, negatively, or uncertainly related, respectively.

In this paper, we use the Audioset sub-tree

of Music [2]. Due to the unitary depth of most child nodes (e.g. Music mood), we will only include the sub-trees of "Musical Instrument" and "Music Genre". A deficiency of using a tree is that inter-category relations cannot be modeled (e.g. "Rock music" and "Guitar"). The triplet-based evaluation methodology can be extended to other graph structures and elaborate ontologies (e.g. WordNet (Miller, 1995)), as well as include intra-category relations (e.g. "Rock music', "Electric guitar", "Viola").

The "Musical Instrument" taxonomy has a maximum depth of 4, encompassing most instrument families, including classical, modern, and non-western instruments. Although it does not separate playing techniques from instruments (e.g., "electric guitar" and "tapping"), omits some instruments (e.g., "viola" from "bowed string instruments") and contains vague concepts ("Musical ensemble"), the taxonomy remains well-defined and free of ambiguous labels.

The "Musical genre" taxonomy has a maximum depth of 3, covering Western music with detailed categorization of contemporary genres (e.g., "Grime music"), as well as folk and non-Western genres. However, it lacks nuance in classical music, only including opera.

3.2 Triplet-based musical knowledge quantification

To curate the music knowledge corpus for LLM evaluation, we leverage the aforementioned subtrees of the Audioset ontology and generate triplets. Specifically, we form triplets of an anchor, a positive and a negative label. The positive and negative labels are defined relative to their semantic similarity with respect to the anchor label. If the anchor is more similar to label 1 than label 2, label 1 is the positive and label 2 is the negative label. This method can encode abstract relationships between labels, including comparisons between non-homogeneous labels (e.g., "happy music", "rock music", "reggae music") but is left for future work as it requires more elaborate ontologies.

We use the distance between the labels based on each tree to quantify their relative similarity. A valid triplet is defined as one where the anchor-positive is less than the anchor-negative distance. After obtaining the valid triplets, we manually inspect them and remove the ones that are ambiguous, vague or too difficult to judge[3].

Finally, we are left with 13633 Genre triplets and 37640 Instrument triplets that will be evaluated separately. Despite the manual inspection, it is important to declare that the dataset is biased toward authors' knowledge of Western music and some triplets might have been erroneously left out.

3.3 Experiments and results

After obtaining the sentence embedding using triplets, cosine similarity will be used to evaluate the relative semantic similarity. Anchor-positive and anchor-negative cosine similarity will be compared and a triplet will be regarded as correct if the first is greater than the second. A thorough analysis of the results is provided in the appendix chapter D. Finally, the accuracy of correct triplets will be calculated and reported.

3.3.1 Prompt sensitivity

Wrapping queried labels in a prompt is useful (Radford et al., 2021) but we are not aware of a thorough analysis of the performance variance concerning different prompts. As a result, we used 20 musically informed prompts. The exact wording of the prompts is provided in appendix C.1. Several words as "music", "recording" or "sound" have been used, to simulate human music captions/descriptions.

The standard deviation reported is relatively high for every case apart from the paraphrased-MiniLM model as presented in table 1. As the prompts do not provide additional information, it can be argued that the models are moderately sensitive to the prompts and "musical" words added can be useful. Lastly, the best model according to model size and performance is paraphrased-ALBERT.

3.3.2 Inability to model negation

Despite the acquired grammatical understanding reported by LLMs, they cannot model negation (e.g. "not rock") (García-Ferrero et al., 2023). To validate if this holds for musical labels, we constructed a separate list of triplets for both "Musical Genre" and "Musical Instruments". For each valid triplet obtained, we extracted unique anchor-positive pairs and introduced a negative label as a negation of the anchor and positive labels. We are left with 3756 and 8284 negative triplets for Genres and Instruments respectively. These were then used alongside 4 negative prompts, listed in appendix C.2.

[2]Visualization: http://www.jordipons.me/apps/audioset/

[3]Removed triplet cases are provided in Appendix table 4

Models	Prompts		Negation	
	Instruments	Genres	Instruments	Genres
mpnet-base	71.3 ± 3.7	76.4 ± 2.3	41.1 ± 3.7	43.2 ± 3.8
distilroberta	62.4 ± 2.4	69.6 ± 2.6	37.2 ± 3.6	42.3 ± 3.4
MiniLM-L12-v2	62.7 ± 2.3	70.9 ± 2.3	33.8 ± 6.5	37.3 ± 6.9
MiniLM-L6	65.8 ± 2.7	70.5 ± 1.6	37.4 ± 5.8	41.4 ± 5.8
Para-albert	69.6 ± 3.2	66.5 ± 1.7	33.4 ± 5.8	35.6 ± 5.7
Para-MiniLM-L3	63.2 ± 2.7	66.9 ± 0.8	29.0 ± 6.7	34.3 ± 5.0

Table 1: Presenting the percentage of correctly inferred triplets for Instruments and Genres respectively. Prompt sensitivity showcased from high standard deviation along 20 prompts. Also, Transformer-based models cannot model negation as the accuracy obtained is worse than random.

Models	Instrument Definitions		Genre Definitions	
	Definition + Label	Definition - Label	Definition - Label	Definition + Label
mpnet-base	83.2 ($\uparrow +11.9$)	72.5 ($\uparrow +1.2$)	84.9 ($\uparrow 8.5$)	72.7 ($\downarrow -3.7$)
distilroberta	75.8 ($\uparrow +13.4$)	73.9 ($\uparrow +11.5$)	71.5 ($\uparrow +1.9$)	69.5 ($\downarrow -0.1$)
MiniLM-L12-v2	81.8 ($\uparrow +19.09$)	72.4 ($\uparrow +9.7$)	79.5 ($\uparrow +8.6$)	70.2 ($\downarrow -0.7$)
MiniLM-L6	80.9 ($\uparrow +15.1$)	72.7 ($\uparrow +6.9$)	79.7 ($\uparrow +9.2$)	69.3 ($\downarrow -1.2$)
Para-albert	79.9 ($\uparrow +10.3$)	68.8 ($\downarrow -0.8$)	80.1 ($\uparrow +13.6$)	74.6 ($\uparrow +8.1$)
Para-MiniLM-L3	81.6 ($\uparrow +18.4$)	67.7 ($\uparrow +4.5$)	76.8 ($\uparrow +9.9$)	70.2 ($\uparrow +3.3$)

Table 2: Results for the experiment showing that models are sensitive towards specific words and cannot properly leverage the context, in the form of a definition. The figures in brackets indicate the difference in accuracy with respect to the experiments with prompts only of table 1.

The performance is worse than random, as shown in table 1, which provides further evidence that LLMs cannot model negation in general and musical terminology. Different prompts lead to considerable differences in accuracy, with the worst performance reported being $\approx 23\%$. This might have potential implications in applications such as captioning, as datasets include negation.

3.3.3 Sensitivity towards the presence of specific words

Using artificially generated definitions of labels instead of generic prompts led to an increased zero-shot image classification accuracy (Pratt et al., 2023). Drawing inspiration from this and leveraging single-sentence definitions provided by Audioset, we evaluate the performance when using the label-free definition and the combination of the label and definition simultaneously.

Excluding the label from the definition leads to a drop in every experiment, meaning that models might be sensitive to labels and not the semantics provided indirectly by the definition. On the other hand, the definition leads to an increment in accuracy in most cases, as shown in table 2.

4 Conclusions and future work

In this paper, we quantified the musical knowledge of six Transformer-based models based on triplet accuracy with musical labels for genres and instruments. We identified three shortcomings: prompt sensitivity, difficulty modeling negation and sensitivity to specific words.

To overcome these shortcomings, we propose using augmentation during training and varying the prompt structures to avoid prompt sensitivity. This approach can utilize definitions to substitute labels with their definitions. To address negation modeling, we suggest multi-task learning that includes tagging negative labels in a caption and maximizing the distance between negative and positive versions of the tags in contrastive losses.

We recommend using lexical databases (e.g. WordNet), which offer more elaborate music concept relationships, instead of using a tree to obtain triplets. We highlight that further filtering needs to be done to form meaningful triplets and produce good-quality evaluation datasets. Lastly, despite reporting increments when definitions are used, further testing is required.

References

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

Marwah Alian and Arafat Awajan. 2020. Factors affecting sentence similarity and paraphrasing identification. *Int. J. Speech Technol.*, 23(4):851–859.

Zaira Hassan Amur, Yew Kwang Hooi, Hina Bhanbhro, Kamran Dahri, and Gul Muhammad Soomro. 2023. Short-text semantic similarity (stss): Techniques, challenges and future perspectives. *Applied Sciences*, 13(6).

Hyunjin Choi, Judong Kim, Seongho Joe, and Youngjune Gwon. 2021. Evaluation of BERT and ALBERT sentence embedding performance on downstream NLP tasks. *CoRR*, abs/2101.10642.

Keunwoo Choi, György Fazekas, Kyunghyun Cho, and Mark Sandler. 2018. The effects of noisy labels on deep convolutional neural networks for music tagging. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2:139–149.

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.

Qingxiu Dong, Xiaojun Wan, and Yue Cao. 2021. Parasci: A large scientific paraphrase dataset for longer paraphrase generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 424–434.

Xingjian Du, Zhesong Yu, Jiaju Lin, Bilei Zhu, and Qiuqiang Kong. 2024. Joint music and language attention models for zero-shot music tagging. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1126–1130. IEEE.

Iker García-Ferrero, Begoña Altuna, Javier Alvez, Itziar Gonzalez-Dios, and German Rigau. 2023. This is not a dataset: A large negation benchmark to challenge large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8596–8615, Singapore. Association for Computational Linguistics.

Josh Gardner, Simon Durand, Daniel Stoller, and Rachel Bittner. 2024. Llark: A multimodal instruction-following language model for music. *Proc. of the International Conference on Machine Learning (ICML)*.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P. W. Ellis. 2022. MuLan: A joint embedding of music audio and natural language. In *International Society for Music Information Retrieval Conference*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *CoRR*, abs/1908.08345.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. 2021. Muscaps: Generating captions for music audio. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. 2022. Contrastive audio-language learning for music. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*.

Ilaria Manco, Benno Weck, Seungheon Doh, Minz Won, Yixiao Zhang, Dmitry Bogdanov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, Elio Quinton, György Fazekas, and Juhan Nam. 2023. The song describer dataset: a corpus of audio captions for music-and-language evaluation. In *Machine Learning for Audio Workshop at NeurIPS 2023*.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41.

Atul Kr. Ojha, A. Seza Doğruöz, Harish Tayyar Madabushi, Giovanni Da San Martino, Sara Rosenthal, and Aiala Rosá, editors. 2024. *Proceedings of the*

18th International Workshop on Semantic Evaluation (SemEval-2024). Association for Computational Linguistics, Mexico City, Mexico.

Sergio Oramas, Luis Espinosa Anke, Francisco Gómez, and Xavier Serra. 2018. Natural language processing for music knowledge discovery. *Journal of New Music Research*, 47:365 – 382.

Sergio Oramas, Mohamed Sordo, Luis Espinosa Anke, and Xavier Serra. 2015. A semantic-based approach for artist similarity. In *International Society for Music Information Retrieval Conference*.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Xiaofei Sun, Yuxian Meng, Xiang Ao, Fei Wu, Tianwei Zhang, Jiwei Li, and Chun Fan. 2022. Sentence Similarity Based on Contexts. *Transactions of the Association for Computational Linguistics*, 10:573–588.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Jan Philip Wahle, Terry Ruas, Tomáš Foltýnek, Norman Meuschke, and Bela Gipp. 2022. Identifying machine-paraphrased plagiarism. In *Information for a Better World: Shaping the Global Future*, pages 393–413, Cham. Springer International Publishing.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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*.

Haoran Xu, Benjamin Van Durme, and Kenton Murray. 2021. Bert, mbert, or bibert? a study on contextualized embeddings for neural machine translation. In *Conference on Empirical Methods in Natural Language Processing*.

Lingling Xu, Haoran Xie, Fu Lee Wang, Xiaohui Tao, Weiming Wang, and Qing Li. 2024. Contrastive sentence representation learning with adaptive false negative cancellation. *Information Fusion*, 102:102065.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32*, pages 5754–5764. Curran Associates, Inc.

Botao Yu, Peiling Lu, Rui Wang, Wei Hu, Xu Tan, Wei Ye, Shikun Zhang, Tao Qin, and Tie-Yan Liu. 2022. Museformer: Transformer with fine- and coarse-grained attention for music generation. In *Advances in Neural Information Processing Systems*, volume 35, pages 1376–1388. Curran Associates, Inc.

A Acknowledgments

The first author is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported jointly by UK Research and Innovation [grant number EP/S022694/1] and Queen Mary University of London.

B Language models used

All the models used are pretrained and then fine-tuned for sentence similarity on several corpora of pairs. Paraphrase models share the same fine-tuning dataset and the same happens for the remaining 4, with an additional 50 million sentence pairs for all-distilroberta-v1. More information can be found in the respective papers, Sentence Transformer [4] package documentation and Hugging Face websites [5].

MPNet unifies the Masked Language Modeling (MLM) and Permuted Language Modeling pretasks, used by BERT (Devlin et al., 2019) and XLNet (Yang et al., 2019) respectively, to train a Transformer backbone. The tokens of the input are permuted, a set of them is masked and the objective is to predict the masked section, while the

[4] https://sbert.net/

[5] https://huggingface.co/

positional information of the full sentence is also known.

DistilBERT (Sanh et al., 2020) is a 40% smaller BERT model that is trained on the same regime as BERT but with an additional loss term. The distillation loss (Hinton et al., 2015) is:

$$L_{ce} = \sum_i t_i * log(s_i) \quad (1)$$

where t_i, s_i is the probability for the predicted tokens of the teacher (BERT) and student (DistilBERT) models respectively. This is used to let the student approximate the target probability distribution of the teacher and therefore, learn from the teacher model.

RoBERTa (Liu et al., 2019) is a model based on BERT with removing next-sentence prediction pretraining, increasing the mini-batch size and altering key hyperparamaters. The analysis of the last are out of the scope for this paper. DistilRoBERTa uses RoBERTa and the distillation process described for DistilBERT.

Instead of approximating the target probability distribution, MiniLM (Wang et al., 2020) proposed to "mimic" the last self-attention module between the student and teacher models. In addition to approximating the attention distribution, this system approximates the relations between the scaled dot-products of queries, keys and value embeddings. Therefore, it also models the second-degree associations between the self-attention embeddings, as well as their distribution.

Finally, ALBERT (Lan et al., 2019) utilizes parameter reduction techniques, as well as swapping the Next Sentence Prediction to Sentence Ordering Prediction. Firstly, Factorized Embedding Parametrization is used to decompose the vocabulary embedding matrix into two small matrices. As a result, the size of the hidden layers is decoupled from the size of the token embeddings. Secondly, Cross-Layer Parameter Sharing relaxes the dependency between memory demands and model depth. Lastly, Sentence Ordering Prediction is focused on predicting the sequence of two sentence segments, while Next Sentence Prediction is used to predict if the pair of sentences is from the same document or not.

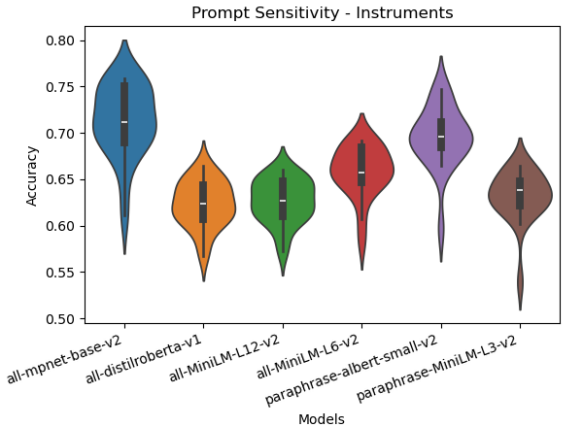


Figure 1: Prompt sensitivity of 6 Transformer-based models with respect to musical instrument terminology.

C Prompts used

C.1 Prompt sensitivity

The prompts used for evaluating the sensitivity towards different musically informed prompts of Transformer-based models are:

1. "*The sound of <label>*"
2. "*Music made with <label>*"
3. "*A <label> track*"
4. "*This is a recording of <label>*"
5. "*A song with <label>*"
6. "*A track with <label> recorded*"
7. "*A music project with <label>*"
8. "*Music made from <label>*"
9. "*Music of <label>*"
10. "*A music recording of <label>*"
11. "*This song is made from <label>*"
12. "*The song has <label>*"
13. "*Music song with <label>*"
14. "*Music song with <label> recorded*"
15. "*Musical sounds from <label>*"
16. "*This song sounds like <label>*"
17. "*This music sounds like <label>*"
18. "*Song with <label> recorded*"
19. "*A <label> music track*"
20. "*Sound of <label>*"

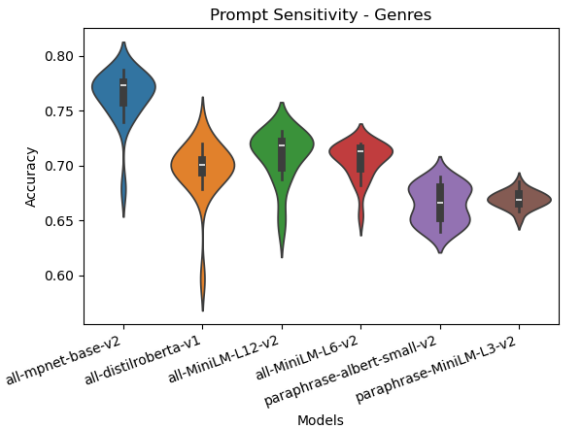


Figure 2: Prompt sensitivity of 6 Transformer-based models with respect to musical genre terminology.

C.2 Negation modeling

The four prompts used to evaluate the inability to model negation:

1. "*No <label>*"
2. "*Not the sound of <label>*"
3. "*Doesn't sound like <label>*"
4. "*Not music from <label>*"

	Instrument Prompts			
Models	**#1**	**#2**	**#3**	**#4**
mpnet-base	45.4	35.9	44.0	39.5
distilroberta	41.4	31.6	39.0	36.7
MiniLM-L12-v2	44.2	30.8	33.6	26.8
MiniLM-L6	46.9	33.2	37.5	32.0
Para-albert	42.7	28.6	28.5	34.0
Para-MiniLM-L3	40.4	23.6	26.8	25.1

(a) Instruments

	Genre Prompts			
Models	**#1**	**#2**	**#3**	**#4**
mpnet-base	49.0	39.5	44.0	40.0
distilroberta	45.6	37.8	45.2	40.1
MiniLM-L12-v2	47.2	35.6	38.7	27.9
MiniLM-L6	49.6	40.7	42.1	33.3
Para-albert	44.8	32.2	29.8	35.7
Para-MiniLM-L3	42.4	32.5	33.0	29.1

(b) Genres

Table 3: Presentation of results for experiment 3.3.2. No model performed on par with the random baseline.

C.3 Examples of removed triplets

As stated in 3.2, there were some triplets of ambiguous quality. We argue that removing these is far more important than building a very big evaluation dataset.

For reference, we present 10 triplets of different ambiguousness levels for each category in table 4.

D Detailed experiment results

D.1 Prompt sensitivity

Generally, prompt sensitivity is evident in every model. The biggest and best model, all-mpnet-base-v2, has the largest and one of the largest variances for instruments (figure 1) and genres respectively (figure 2).

Paraphrase-MiniLM-L3-v2 had the smallest variance for genres, at the expense of a lower accuracy. This might be due to the different distillation process chosen. If an application demands robustness towards prompt sensitivity, that would be the best choice.

Apart from all-mpnet-base-v2, every model had approximately the same variance when the outliers were discarded, as can be seen in figure 1.

D.2 Negation modeling

By far the worst deficiency found is the inability of Transformer-based models to model negation. These failed to surpass random choice in every experiment, while altering the prompt led to a significant decrease in accuracy, up to $\approx 20\%$. This is presented in table 3a.

This result can have large implications on developing or evaluating captioning systems, as datasets (Agostinelli et al., 2023; Manco et al., 2023) contain negation and following these results, can lead to erroneous inference. Also, joint audio-text models, also known as two-tower systems, can be negatively impacted. Further testing is required in the future.

Instruments		
Anchor	Positive	Negative
Musical instrument	Plucked string instrument	Mandolin
Cowbell	Accordion	Flute
Guitar	French horn	Timpani
Electric guitar	Hammond organ	Rhodes piano
Bass guitar	Brass Instrument	Alto saxophone
Tapping (guitar technique)	French horn	Electric piano
Sitar	Cymbal	Rimshot
Keyboard (musical)	Cowbell	Acoustic guitar
Piano	Didgeridoo	Cello
Organ	Trombone	Timpani
Genres		
Anchor	Positive	Negative
Music genre	Rhythm and blues	Swing music
Pop music	Jazz	Swing music
Hip hop music	Classical music	Drum and bass
Rock music	Independent music	Grime music
Heavy metal	Electronic music	Oldschool jungle
Progressive rock	Chant	Oldschool jungle
Reggae	Music of Asia	Cumbia
Jazz	New-age music	Heavy metal
Kuduro	Music for children	Grunge
Funk carioca	Christian music	Electronica

Table 4: Table with examples of removed triplets. The filtering criterion is based on the ambiguity or relative difficulty in determining whether the anchor is more similar to the positive or negative label.

FUTGA: Towards Fine-grained Music Understanding through Temporally-enhanced Generative Augmentation

Junda Wu[1] Zachary Novack[1] Amit Namburi[1] Jiaheng Dai[1] Hao-Wen Dong[1]
Zhouhang Xie[1] Carol Chen[2] Julian McAuley[1]

[1] Computer Science and Engineering, UC San Diego [2] Computer Science Department, UC Los Angeles

Abstract

We propose FUTGA, a model equipped with fined-grained music understanding capabilities through learning from generative augmentation with temporal compositions. We leverage existing music caption datasets and large language models (LLMs) to synthesize fine-grained music captions with structural descriptions and time boundaries for full-length songs. Augmented by the proposed synthetic dataset, FUTGA is enabled to identify the music's temporal changes at key transition points and their musical functions, as well as generate detailed descriptions for each music segment. We further introduce a full-length music caption dataset generated by FUTGA, as the augmentation of the MusicCaps and the Song Describer datasets. The experiments demonstrate the better quality of the generated captions, which capture the time boundaries of long-form music. Generated temporal-aware music descriptions are illustrated in our demonstration https://namburiamit.github.io/futga.github.io/.

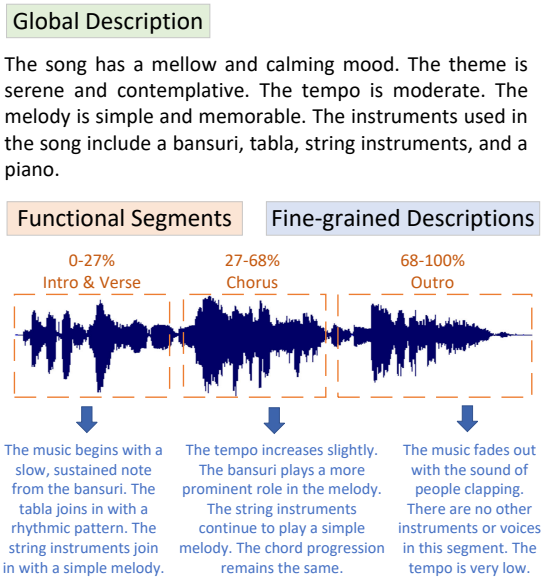


Figure 1: **Overview of FUTGA's capabilities.** Given a long-form audio example, FUTGA is able to provide time-located captions by automatically detecting functional segment boundaries, as well as global captions.

1 Introduction

Natural language music understanding, which extracts music information and generates detailed music captions, is a fundamental task within the MIR community, beneficial for a series of applications including music generation (Copet et al., 2024; Chen et al., 2024; Novack et al., 2024; Melechovsky et al., 2023), editing (Wang et al., 2023; Zhang et al., 2024), question-answering (Deng et al., 2023; Gao et al., 2022), and retrieval (Doh et al., 2023b; Wu et al., 2023; Bhargav et al., 2023). Recent developments in music foundation models (Gardner et al., 2023; Hussain et al., 2023; Tang et al., 2023; Liu et al., 2024) enable free-form music prompts and multitasking. These foundation models are developed based on pre-trained large language models (LLMs) and aligned with the music modality. Although LLM-powered music understanding models can leverage the abundant pre-trained music knowledge in caption generation, the success of modality alignment still requires a large amount of high-quality music caption data.

Restricted by the current music captioning paradigm, available music caption datasets are limited to two major challenges: (1) Conventional music captions focus only on the global description of a (potentially long) music clip, which cannot efficiently capture a piece of music's fine-grained characteristics nor differentiate it from other music within-genre songs. (2) Key structural information, such as time boundaries of functional music segments and time-aware musical changes, is mostly neglected in traditional music understanding and hard to retrieve due to the limitation in the length of music clips.

To address the limitations, we propose **FUTGA**, a generative music understanding model trained with time-aware music caption data and cali-

brated with Music Information Retrieval (MIR) features. We first augment the MusicCaps dataset (Agostinelli et al., 2023) by mixing music clips together into synthetic full-length songs. The corresponding music captions are composed with original short music captions as individual segment descriptions, which are also tagged with temporal segmentation information. To enable more realistic full-length music captioning, we further leverage a text-only LLM for the augmentation of the global music caption, musical changes between segments (*e.g.*, increase of volume, slowing down the tempo, introducing new instruments, etc.), and functional tags of the segments (*e.g.*, intro, verse, chorus, etc.), by paraphrasing and summarizing the template-based captions.

Inspired by existing Large Audio-Language Models (LALMs), we use the open-source SALMONN model (Tang et al., 2023) as the backbone and fine-tune the model with our developed synthetic full-length music caption data. Using our synthetic data augmentation, FUTGA is able to identify key transition points in musical changes and segment full-length songs according to their musical functions. For example, in Figure 1, we illustrate FUTGA's capacities as a novel form of music captioning. Given a song in full length, FUTGA can generate a global caption that summarizes the whole song's characteristics before identifying the music structure with time segments. Following the flow of music structures, FUTGA can further describe each music clip and musical changes between consecutive music clips. In addition, we also discover that the fine-tuned SALMONN model demonstrates a great instruction-following capacity to generate fine-grained music captions conditioned on given time boundaries and MIR features. By injecting the ground-truth information into the instruction prompt, we can accurately guide the model to generate fine-grained music captions corresponding to the time segments. With the final version of FUTGA, we propose automatically annotating the full-length songs in two existing datasets MusicCaps (Agostinelli et al., 2023) and Song Describer (Manco et al., 2023).

2 Temporally-enhanced Generative Augmentation

In this section, we introduce our proposed temporally-enhanced generative augmentation. Due to the limitation of existing music caption datasets, music captioning and understanding models can only generate global music descriptions for short music clips (Manco et al., 2023; Agostinelli et al., 2023; Doh et al., 2023a). To address this limitation, we propose the augmentation of synthetic music and caption composition, which empowers music understanding models with capacities of time-aware music segmentation and fine-grained music description generation.

For each sampled set of music clips C_k, the corresponding music caption set T_k and the clip length information L_k are interleaved and composed by the template,

$$\tilde{X}_k = \left\{ \left(\frac{l_{k,j-1}}{\sum_i l_{k,i}}, \frac{l_{k,j}}{\sum_i l_{k,i}}, t_{k,j} \right) \right\}_{j=1}^{n},$$

in which the specific original time-boundaries L_k are transformed into relative time-boundaries, which are always between $0 - 100\%$ ($l_{k,0} \equiv 0$). We use a relative time-boundary representation approach to minimize training bias towards specific numbers of music lengths in our model. In addition, relative time-boundary representation enables the model's ability to comprehend music of varying lengths, thereby improving the model's generalizability. By generating these relative time boundaries, our generative music understanding model gains a better awareness of the music's overall progression, which further enhances the model's temporal understanding of music.

To further augment rich MIR features in music captions, we further propose to use a text-only large language model (LLM) to augment the template-based caption $\tilde{X}_k$ with natural language descriptions, in which additional information, such as global captions, musical changes, and music structures, can be automatically extracted from the LLM. Since LLMs are pre-trained with abundant domain knowledge including music analysis (Tang et al., 2023; Gardner et al., 2023) and music information retrieval (Tang et al., 2023; Hussain et al., 2023), with enough context provided, such LLMs can accurately extract music information via language-based summarization and reasoning (Doh et al., 2023a). Inspired by LLMs' capacities in language reasoning, we propose to paraphrase and augment additional music information with instructions as follows:

Context: *Music Analysis* $\{\tilde{X}_k\}$. *This is a music analysis of a song. Note that the numbers indicate the time-boundaries of functional segments in this*

MusicCaps: The low quality recording features footsteps, birds chirping, distant traffic and wind sound effects. At the very end, there is a suspenseful, low, widely spread pad chord. It sounds like it is a sound for some kind of video.
Song Describer: A rock song with a French speaking male voice. It has the typical feeling of political/social rock song with a really simple guitar, bass and drum skeleton.
LP-MusicCaps: The low quality recording features a drum solo that consists of a double pedal kick, punchy snare, shimmering cymbals and boomy toms. It sounds energetic and manic, thanks to that kick pattern. The recording is mono and noisy.
FUTGA: The song exhibits a diverse range of musical styles and instrumentation. Overall, the mood is upbeat and energetic, with a theme of youthfulness and optimism. The tempo is moderate to fast, with a catchy melody and a simple chord progression. **0-27%: Intro and Verse** - The song begins with a haunting intro featuring syncopated rhythms and atmospheric chords. The verse introduces the main melody, which is characterized by a catchy hook and a simple chord progression. **27-58%: Chorus and Bridge** - The chorus features a soaring melody and energetic vocals. The bridge introduces a new melody and chord progression, which contrasts with the verse and chorus. **58-79%: Rock Section** - The song transitions into a rock section with a distorted electric guitar and a heavy drum beat. The section has a groovy feel and could be used in the soundtrack of a high school drama TV series. **79-100%: Acoustic Ballad** - The song concludes with an acoustic ballad, sung in a melancholic manner. The melody is simple and the chord progression is repetitive.

Table 1: A comparison example of captions generated or annotated by MusicCaps (Agostinelli et al., 2023), Song Describer (Manco et al., 2023), LP-MusicCaps (Doh et al., 2023a) and FUTGA.

Dataset	# Caption	# Segment	Tokens	Vocab.	# Inst.	# Genre	# Mood
MusicCaps	6k	–	48.9 ± 17.3	6,144	75	267	146
Song Describer	1k	–	21.7 ± 12.4	2,859	39	152	122
LP-MusicCaps	542k	–	45.3 ± 28.0	1,686	65	239	151
FUTGA	7k	4.32	472.419 ± 88.5	3,537	64	187	128

Table 2: Statistics summarization of generated or annotated music captions of baselines and FUTGA.

song.

Paraphrase: *Paraphrase the music analysis to make it sound like a coherent song, instead of a remix. Additionally, remove any mention of sound quality.*

Global Caption: *Start with a general description of the song focusing on subjectivity.*

Musical Change: *Describe the song in detail and explain transitions between parts of the song.*

Music Structure: *Remember to indicate the temporal annotations and music structures when talking about a specific part of the song.*

3 Dataset Creation: FUTGA

Based on the final version of our proposed music captioning model, we automatically generate music captions for whole songs between 2 minutes and 5 minutes in MusicCaps (Agostinelli et al., 2023) and Song Describer (Manco et al., 2023). During inference time, we set the repetition penalty as 1.5 to prevent repetitive descriptions of the same music segments. In addition, we also set the beam search number to 10 to find the statistically best captions. We allow a maximum of 2048 tokens to be generated from FUTGA.

As demonstrated in the comparison example in Table 1, FUTGA provides more fine-grained music understanding descriptions with time boundaries indicating music segments, for which the average segment number and the number of musical changes are reported in Table 2. In addition, we can observe relatively longer global captions with more details, which is also verified by the data statistics in Table 2.

In terms of music caption diversity, we first show that our captions have significantly larger numbers of tokens and vocabulary size, compared to existing music caption datasets. Second, our dataset still maintains good diversity in terms of unique genre, instrument, and music mood vocabularies, which are comparable to human or GPT-3.5 annotations. Thus, the FUTGA dataset can serve to augment existing music captioning models with strong temporal reasoning abilities without harming the model's generalizability, which will be further evaluated in our evaluation section.

	MusicCaps						Song Describer					
Model	B1	B2	B3	M	R	B-S	B1	B2	B3	M	R	B-S
LP-MusicCaps	19.77	6.70	2.17	12.88	13.03	84.51	1.68	0.71	0.27	7.68	2.76	79.62
FUTGA (complete)	9.21	4.18	1.97	**20.85**	11.96	82.62	4.58	1.72	0.61	12.82	6.90	81.21
FUTGA (global)	**26.46**	**10.93**	**4.66**	18.60	**17.40**	**86.48**	**14.23**	**5.04**	**1.75**	**15.04**	**11.67**	**85.42**

Table 3: Comparison results of caption generation for LP-MusicCaps and FUTGA.

4 Experiments and Results

We obtain 5K synthetic training samples by prompting the GEMMA-7B model (Team et al., 2024) with the template-based caption $\tilde{X}_k$ and the designed instructions. Then we adopt LoRA (Hu et al., 2021) instruction finetuing of the SALMONN-7B (Tang et al., 2023) backbone model for 100 epochs and the learning rate of 1e-5, with 2 NVIDIA RTX A6000 GPUs with 48GB each. We use the bfloat16 type for training with the batch size set to 4 and gradient accumulation steps to 8.

We first evaluate the generated data samples' quality by comparing them to existing human annotation datasets, MusicCaps (Agostinelli et al., 2023) and Song Describer (Manco et al., 2023). We follow the previous works (Doh et al., 2023a; Manco et al., 2023) and report the metrics, BLEU (B), METEOR (M), ROUGE (R), and BERT-score (B-S), in Table 3. Since our captions are formally different from original music captions, we report the evaluation metrics for the global and the complete captions in our dataset separately. For a fair comparison, we adopt the zero-shot performance of LP-MusicCaps in (Doh et al., 2023a), since our model is only trained on the synthetic dataset and Harmonixset.

Based on the results in Table 3 on MusicCaps, we observe that the global captions generated from our model consistently show higher quality than the zero-shot results of LP-MusicCaps, which demonstrates that by capturing more details from longer songs, we can obtain more accurate descriptions of the music. In addition, comparing FUTGA and LP-MusicCaps on Song Describer, which is the out-of-domain dataset for both methods, FUTGA shows a significantly larger improvement in the generation results, which demonstrates the model's better capacities in generalizability.

However, the complete music captions generated from FUTGA show relatively inferior performance on MusicCaps, which is mainly due to the different forms of music captions. Since FUTGA focuses on the temporal reasoning of a whole song, the time segment information and musical changes are completely new to both the original MusicCaps captions. Whereas, LP-MusicCaps is directly augmented from MusicCaps, which makes their captions formally more similar. Such observations can motivate future works to explore more fine-grained and complex music caption forms in terms of evaluating the model's generation capacities.

5 Conclusion

In this work, we propose a temporally-enhanced music caption augmentation method through generative large language models. By bootstrapping existing music captions with time boundary tags, MIR features, and musical changes, we fine-tune the pre-trained music understanding model SALMONN-7B, where we observe emerging music segmentation capacities and enable instruction prompting to guide the generation with ground-truth time segments. We use the fine-tuned model to re-annotate the existing MusicCaps and Song Describer datasets with full-length songs. The generated captions are shown to be more fine-grained and beneficial for various downstream tasks.

For future works, since our model is the first to enable end-to-end full-length song captioning with significantly longer context provided (10 times more than conventional music captions), we are motivated to further develop a long-context-based CLAP model, which can enable more complex and longer music retrieval tasks. In addition, with more fine-grained details provided by our captions, we propose to further use such captions for more complex music understanding tasks, including music question-answering and whole-song generation.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Samarth Bhargav, Anne Schuth, and Claudia Hauff. 2023. When the music stops: Tip-of-the-tongue retrieval for music. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2506–2510.

Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2024. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1206–1210. IEEE.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2024. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36.

Zihao Deng, Yinghao Ma, Yudong Liu, Rongchen Guo, Ge Zhang, Wenhu Chen, Wenhao Huang, and Emmanouil Benetos. 2023. Musilingo: Bridging music and text with pre-trained language models for music captioning and query response. *arXiv preprint arXiv:2309.08730*.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023a. Lp-musiccaps: Llm-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*.

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)*, pages 1–5. IEEE.

Wenhao Gao, Xiaobing Li, Cong Jin, and Yun Tie. 2022. Music question answering: Cognize and perceive music. In *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE.

Josh Gardner, Simon Durand, Daniel Stoller, and Rachel M Bittner. 2023. Llark: A multimodal foundation model for music. *arXiv preprint arXiv:2310.07160*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Atin Sakkeer Hussain, Shansong Liu, Chenshuo Sun, and Ying Shan. 2023. M Θ{2} ugen: Multi-modal music understanding and generation with the power of large language models. *arXiv preprint arXiv:2311.11255*.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2024. Music understanding llama: Advancing text-to-music generation with question answering and captioning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 286–290. IEEE.

Ilaria Manco, Benno Weck, Seungheon Doh, Minz Won, Yixiao Zhang, Dmitry Bodganov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, et al. 2023. The song describer dataset: A corpus of audio captions for music-and-language evaluation. *arXiv preprint arXiv:2311.10057*.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2023. Mustango: Toward controllable text-to-music generation. *arXiv preprint arXiv:2311.08355*.

Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J Bryan. 2024. Ditto: Diffusion inference-time t-optimization for music generation. *arXiv preprint arXiv:2401.12179*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, et al. 2023. Audit: Audio editing by following instructions with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:71340–71357.

Shangda Wu, Dingyao Yu, Xu Tan, and Maosong Sun. 2023. Clamp: Contrastive language-music pre-training for cross-modal symbolic music information retrieval. *arXiv preprint arXiv:2304.11029*.

Yixiao Zhang, Yukara Ikemiya, Gus Xia, Naoki Murata, Marco Martínez, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. 2024. Musicmagus: Zero-shot text-to-music editing via diffusion models. *arXiv preprint arXiv:2402.06178*.

The Interpretation Gap in Text-to-Music Generation Models

Yongyi Zang[*]
Independent Researcher
zyy0116@gmail.com

Yixiao Zhang[*]
C4DM, Queen Mary University of London
yixiao.zhang@qmul.ac.uk

Abstract

Large-scale text-to-music generation models have significantly enhanced music creation capabilities, offering unprecedented creative freedom. However, their ability to collaborate effectively with human musicians remains limited. In this paper, we propose a framework to describe the musical interaction process, which includes expression, interpretation, and execution of controls. Following this framework, we argue that the primary gap between existing text-to-music models and musicians lies in the interpretation stage, where models lack the ability to interpret controls from musicians. We also propose two strategies to address this gap and call on the music information retrieval community to tackle the interpretation challenge to improve human-AI musical collaboration.

1 Introduction

In recent years, the field of human-AI music co-creation has experienced significant advancements (Huang et al., 2020; Zhang et al., 2021; Rau et al., 2022; Bougueng Tchemeube et al., 2022). The advent of large-scale text-to-music generation models has played a crucial role in this progress, enabling generating music with good sonic quality and well-defined musical structures (Copet et al., 2024; Evans et al., 2024; Agostinelli et al., 2023).

A primary focus of recent research has been to enhance these models through the incorporation of control signals (Lin et al., 2023; Tal et al., 2024; Wu et al., 2024; Lin et al., 2024; Nistal et al., 2024). This has led to significant success in manipulating dynamics, melody, and chord progressions in generated music contents. While precision in following these control signals can still be improved, these developments represent substantial progress.

Although extensive efforts are made to allow these models follow control signals precisely, misalignments between musicians' intent and model output still exist, making effective collaboration with musicians challenging (Yakura and Goto, 2023; Newman et al., 2023; Ronchini et al., 2024; Majumder et al., 2024). In practice, we observe that musicians' control signals require interpretation before execution, and this process is often overlooked in current music information retrieval research. This oversight may hinder the practical applicability of these models in real-world musical settings. Figure 1 illustrates this issue through a single-round interaction among musician A and either musician B or a model. In this interaction, the control signals expressed by musician A are successfully interpreted by musician B before B generates the musical outputs. In contrast, the model fails to interpret these signals due to the neglected interpretation process in current text-to-music generation models.

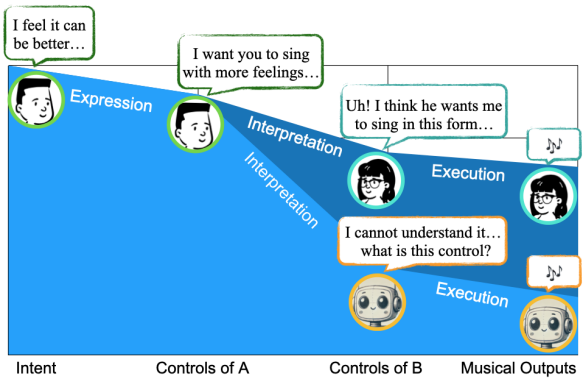


Figure 1: The comparison between human-human and human-AI interaction processes. We observe that the gap exists at both the interpretation stage and the execution stage, while the interpretation stage is often overlooked by current research.

In this paper, our contribution is threefold:

1. We propose a framework for the musical interaction process, consisting of three stages: expression, interpretation, and execution of control.
2. Our literature review identifies a communication

[*]Both authors contributed equally.

gap in current models, which often fail to interpret controls in a way that aligns with human musicians' natural communication methods.

3. To address this gap, we propose two approaches: directly learning from human interpretation data or leveraging a strong prior understanding of human interpretation, such as that found in large language models (LLMs).

2 Interpretation of Controls

To begin with, we propose a general framework that conceptualizes the musical interaction procedure in three stages: the *expression*, *interpretation*, and *execution* of controls, as shown in Figure 2.

Figure 2: The proposed model that describes musical interaction process.

In an interaction between parties A and B:

- **Expression**: A's intent is mapped to *Controls A*;
- **Interpretation**: B interprets *Controls A*, resulting in *Controls B*;
- **Execution**: B executes *Controls B*, producing the final musical output.

Table 1 provides several examples illustrating this framework. The framework encompasses both solo and multi-party musical interactions, with the interpretation stage becoming explicit in multi-party scenarios. Successful realization of the original intent hinges on effective mapping across all three stages of the process.

In this section, we examine the musical communication process following this framework. We observe that musical interactions often involve varying degrees of ambiguity in control expression, and skilled musicians can effectively interpret and execute these ambiguous instructions. In contrast, current text-to-music generation models struggle with this ambiguity, and can only understand highly semantical or highly precise instructions.

2.1 Musicians' Interpretation of Controls

Musicians communicate through varying levels of ambiguity (Bishop, 2018). The most precise instructions often point to only one outcome (e.g., "Turn the bass 3 dB up") while the most abstract ones require much creative interpretation (e.g., "I want a *moody* synth"). Most communications, however, lie between these two extremes.

Consider this example of a producer addressing a vocalist: [1] "I want to try one where you just start this *chorus very soft*, and in that first phrase, like *[inaudible]*. You know what I mean? *(Sing to demonstrate)* just like, *get crazy* with it. Let's *start quieter* ... or *softer*, or, *babier*. Just try it." This example showcases a wide range of communication types, from highly semantic descriptions (e.g., "very soft," "get crazy," "quieter," "softer," "babier") to performative instructions (e.g., "(Sing to demonstrate)"), and others that fall somewhere in between, requiring interpretation (e.g., "chorus," "first phrase," "[inaudible]").

Human musicians excel at interpreting musical instructions with varying ambiguity, a skill known as "musical taste" or "musicianship" (Sloboda, 1986). This ability enables jazz musicians to adapt improvisations (Berliner, 2009), film composers to modify scores for evolving narratives (Cooke, 2008), and orchestral conductors to guide an ensemble through gestures (Bishop et al., 2019) and verbal cues. This skill, which develops with experience (Lehmann et al., 2007), involves intuitive understanding of musical context, style, and intent (Meyer, 2008), allowing musicians to transform ambiguous directions into coherent expressions (Daniel et al., 2006).

2.2 Models' Interpretation of Controls

While human musicians excel at interpreting ambiguous instructions, current music generation models struggle with this task. Traditional approaches to control often rely on disentangling representations in latent space (Luo et al., 2019; Wang et al., 2020). For music generative models, control mechanisms are typically implemented through various strategies. Some models integrate controls during initial large-scale pre-training, such as Mustango (Melechovsky et al., 2024) and MusicGen (Copet et al., 2024). Others employ post-training model augmentation, exemplified by Cocomulla (Lin et al., 2023), AIRGen (Lin et al., 2024), and Music ControlNet (Wu et al., 2024). Additionally, some approaches combine both stages' efforts, as seen in MusicMagus (Zhang et al., 2024b),

[1] Billie Eilish In Studio Making Album "When We All Fall Asleep, Where Do We Go?, https://www.youtube.com/watch?v=Sp-eNvKV0to

Cases	Intent		Controls A		Controls B		Outputs
Solo Interactions							
Pianist	Light touch	→	Reduce finger force	→	N/A	→	Piano audio
Experienced Producer	Spacious sound	→	Reverb, cut lows	→	N/A	→	Natural result
Novice Producer	Spacious sound	→	Only adding reverb	→	N/A	→	Unnatural result
Composer	Modulate key	→	Write transition	→	N/A	→	Score
Experienced Guitarist	Emphasizing a chord	→	Use complex fingering	→	N/A	→	Clean strum sound
Novice Guitarist	Emphasizing a chord	→	Use complex fingering	→	N/A	→	Muffled strum sound
Multi-Party Interactions							
Producer & Experienced Vocalist	Emotive singing	→	"More feelings"	→	More dynamics & articulation	→	Emotional vocal track
Producer & Novice Vocalist	Emotive singing	→	"More feelings"	→	Sing closer to microphone	→	Unnatural vocal track
Experienced Rock Band	Guitar solo	→	Gesture	→	Drums and bass play fill; vocalist stop singing	→	Solo section
Novice Rock Band	Guitar solo	→	Gesture	→	Everyone ignores the guitarist	→	Solo fights with vocal, creating cacophony
Conductor & Orchestra	Crescendo	→	Rising arms	→	Gradually increasing dynamics	→	Balanced crescendo
DJ & Crowd	Build energy	→	Throwing hands up in the air	→	Crowd thinks it's peak	→	Premature movement

Table 1: Examples of solo and multi-party musical interactions.

Instruct-MusicGen (Zhang et al., 2024a), and Chat-Musician (Yuan et al., 2024a). Despite these advancements in control capabilities, current models still fall short of matching human-level interpretation of nuanced musical instructions.

We posit that the challenge lies not in control implementation methods, but in the nature of the controls themselves. Table 2 summarizes the controls offered by current models, typically either highly semantic (e.g., text descriptions) or highly specific (e.g., chords, melodies). These models struggle with both: for semantic inputs, they mainly interpret at keyword-level rather than understanding natural language (Wu et al., 2023), failing with concepts like negation and temporal order (Agostinelli et al., 2023; Yuan et al., 2024b); for specific inputs, they struggle with precise execution (Zhang et al., 2024a). When prompts combine semantic and specific instructions, models often fail to interpret the former and fail to execute the latter. The lack of support for other modalities, such as visual cues, also makes effective interpretation more difficult.

While resolving all these challenges is crucial, current research primarily focuses on improving execution ability, such as audio quality, while largely overlooking the interpretation stage. This oversight creates a significant gap in human-AI musical collaboration. Musicians are forced to adapt to the constrained and unnatural controls offered by these models, rather than the models adapting to musicians' natural communication methods. We posit that this mismatch is a key factor in the limited adoption of these otherwise highly capable models by musicians in practice.

3 Potential Solutions to Improve Interpretation of Controls

Addressing the interpretation gap between musicians and models is challenging due to the complex, multi-modal nature of musician communication, which includes visual cues, textual prompts, vocalizations, and musical references. No existing data sources comprehensively capture all modalities of music interactions, and creating such a dataset would be resource-intensive. Thus, we must approach the problem of learning interpretation under resource constraints. Given these limitations, two potential solutions emerge: directly learning from many aspects of human interpretation data, or leveraging a strong prior understanding of human interpretation, such as that encapsulated in large language models (LLMs). In the following sections, we explore these two avenues for enhancing AI models' ability to interpret musical controls.

3.1 Directly Learn from Human Interpretation Data

Previous research has explored many aspects of musical perception and interpretation, including auditory perception (Ananthabhotla et al., 2019; Wright and Välimäki, 2020; Manocha et al., 2020), emotion (Yang and Chen, 2012; Dash and Agres, 2023), song and artist similarity (Knees and Schedl, 2013; Allik et al., 2018), music discussions (Hauger et al., 2013), recommendation systems (Bertin-Mahieux et al., 2011), and non-verbal communications, such

Model	Semantic controls	Precise controls
Integrated Controls in Foundation Models		
Mustango (Melechovsky et al., 2024)	Text description, metadata	-
MusicGen (Copet et al., 2024)	Text description	melody spectrogram
Diff-A-Riff (Nistal et al., 2024)	Text description	Music audio mixture
Jen-1 Composer (Yao et al., 2023)	Text description	Other instrument tracks
GMSDI (Postolache et al., 2024)	Instrument name	Other instrument tracks
Control Enhancement Modules		
Coco-mulla (Lin et al., 2023)	Text description	Drum track, chord, melody
AIRGen (Lin et al., 2024)	Text description	Drum track, chord, melody
JASCO (Tal et al., 2024)	Text description	Drum track, chord, melody
Music ControlNet (Wu et al., 2024)	Text description	Dynamic, melody, rhythm
Jen-1 DreamStyler (Chen et al., 2024)	Text description	Reference music audio
Music Editing Methods		
MusicMagus (Zhang et al., 2024b)	Text swapping	Music audio mixture
InstructME (Han et al., 2023)	Edit instruction	Music audio mixture
Instruct-MusicGen (Zhang et al., 2024a)	Edit instruction	Music audio mixture
Loop Copilot (Zhang et al., 2023)	Edit instruction	Conversational context (music audio, text)
M^2UGen (Hussain et al., 2023)	Edit instruction	Conversational context (music audio, text)
ChatMusician (Yuan et al., 2024a)	Edit instruction	Conversational context (symbolic music, text)

Table 2: List of representative text-to-music generation models with extra controls. Most controls can be classified into high-level semantic controls and low-level signal-level controls, while the exploration of intermediate-level musicians' communication controls are limited.

as gesture and dance movements (Gillian, 2012; Fan et al., 2011). These studies often rely on crowd-sourced evaluations or public data, achieving good interpretations that can serve as control signals, as demonstrated by Huang et al. (2024) in music generation from dance movements.

Learning directly from these diverse sources require combining them into cohensive controls, which may be achieved through pseudo-description generation, an approach that has shown promise in music captioning (Mei et al., 2024; Doh et al., 2023) and understanding (Liu et al., 2024).

3.2 LLMs for Musical Interpretation

LLMs' robust language understanding enables the decomposition of user queries into specialized tasks, an approach pioneered by HuggingGPT (Shen et al., 2024). This method has inspired audio domain projects such as Loop Copilot (Zhang et al., 2023), WavJourney (Liu et al., 2023), WavCraft (Liang et al., 2024), and MusicAgent (Yu et al., 2023). Jiang et al. (2024) explores synthesizing natural language from control parameters for model training.

However, user studies (Gianet et al., 2024; Newman et al., 2023; Ronchini et al., 2024; Zhang et al., 2023) reveal that professional musicians often experience misalignment between model interpretations and their intentions, primarily due to LLMs' lack of domain-specific musical knowledge (Li et al., 2024). Research in other domains indicates that simply integrating domain knowledge can significantly enhance LLMs' capabilities (Lee et al., 2024). Consequently, we posit that by collecting domain knowledge and natural music conversations incorporating this knowledge, we could effectively boost LLMs' ability to execute music tasks. Furthermore, these enhanced LLMs could potentially generate synthetic training data for developing more compact interpretation models.

4 Conclusion

We identify a critical gap in text-to-music generation models: their inability to effectively interpret musicians' controls. We propose a three-stage framework for musical interaction: expression, interpretation, and execution, and highlight how current AI models often struggle with the crucial interpretation stage. To address this gap, we suggest two potential solutions: directly learning from various sources of human interpretation data and leveraging large language models for musical interpretation. We call on the MIR community to prioritize research in this area, as improving the interpretation capabilities is crucial for their integration into creative workflows and for realizing their full potential as collaborative tools for musicians.

Ethics Statement

Our work includes YouTube video transcript excerpts demonstrating artists' creative processes, used solely to illustrate our proposed framework. We thank these amazing artists for sharing their creative processes. All copyrights remain with the original video owners, and excerpts are included for research purposes only.

We acknowledge that musical communications and interpretations encapsulate diverse musicianship, tastes, and cultural nuances. While some aspects of musical communications may be universal, they are often influenced by social culture and individual experiences. We encourage the community to be mindful of this diversity when modeling musical interpretations, as capturing these nuances can enhance the music creation process with generative models.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Alo Allik, Florian Thalmann, and Mark Sandler. 2018. Musiclynx: Exploring music through artist similarity graphs. In *Companion Proceedings of the The Web Conference 2018*, pages 167–170.

Ishwarya Ananthabhotla, Sebastian Ewert, and Joseph A Paradiso. 2019. Towards a perceptual loss: Using a neural network codec approximation as a loss for generative audio models. In *Proceedings of the 27th acm international conference on multimedia*, pages 1518–1525.

Paul F Berliner. 2009. *Thinking in jazz: The infinite art of improvisation*. University of Chicago Press.

Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset.

Laura Bishop. 2018. Collaborative musical creativity: How ensembles coordinate spontaneity. *Frontiers in psychology*, 9:1285.

Laura Bishop, Carlos Cancino-Chacón, and Werner Goebl. 2019. Moving to communicate, moving to interact: Patterns of body motion in musical duo performance. *Music Perception: An Interdisciplinary Journal*, 37(1):1–25.

Renaud Bougueng Tchemeube, Jeffrey John Ens, and Philippe Pasquier. 2022. Calliope: A co-creative interface for multi-track music generation. In *Proceedings of the 14th Conference on Creativity and Cognition*, pages 608–611.

Boyu Chen, Peike Li, Yao Yao, and Alex Wang. 2024. Jen-1 dreamstyler: Customized musical concept learning via pivotal parameters tuning. *arXiv preprint arXiv:2406.12292*.

Mervyn Cooke. 2008. *A history of film music*. Cambridge University Press.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2024. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36.

J Daniel et al. 2006. This is your brain on music: The science of a human obsession.

Adyasha Dash and Kathleen Agres. 2023. Ai-based affective music generation systems: A review of methods and challenges. *ACM Computing Surveys*.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023. Lp-musiccaps: Llm-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*.

Zach Evans, CJ Carr, Josiah Taylor, Scott H Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. In *Forty-first International Conference on Machine Learning*.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. 2024. Long-form music generation with latent diffusion. *arXiv preprint arXiv:2404.10301*.

Rukun Fan, Songhua Xu, and Weidong Geng. 2011. Example-based automatic music-driven conventional dance motion synthesis. *IEEE transactions on visualization and computer graphics*, 18(3):501–515.

Eric Tron Gianet, Luigi Di Caro, and Amon Rapp. 2024. Music composition as a lens for understanding human-ai collaboration.

Nicholas Edward Gillian. 2012. *Gesture recognition for musician computer interaction*. Ph.D. thesis, Citeseer.

Bing Han, Junyu Dai, Xuchen Song, Weituo Hao, Xinyan He, Dong Guo, Jitong Chen, Yuxuan Wang, and Yanmin Qian. 2023. Instructme: An instruction guided music edit and remix framework with latent diffusion models. *arXiv preprint arXiv:2308.14360*.

David Hauger, Markus Schedl, Andrej Košir, and Marko Tkalcic. 2013. The million musical tweets dataset: What can we learn from microblogs. In *Proc. ISMIR*, pages 189–194.

Cheng-Zhi Anna Huang, Hendrik Vincent Koops, Ed Newton-Rex, Monica Dinculescu, and Carrie J Cai. 2020. Ai song contest: Human-ai co-creation in songwriting. *arXiv preprint arXiv:2010.05388*.

Jiang Huang, Xianglin Huang, Lifang Yang, and Zhulin Tao. 2024. Dance-conditioned artistic music generation by creative-gan. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 107(5):836–844.

Atin Sakkeer Hussain, Shansong Liu, Chenshuo Sun, and Ying Shan. 2023. M^2ugen: Multi-modal music understanding and generation with the power of large language models. *arXiv preprint arXiv:2311.11255*.

Xilin Jiang, Cong Han, Yinghao Aaron Li, and Nima Mesgarani. 2024. Listen, chat, and edit: Text-guided soundscape modification for enhanced auditory experience. *arXiv preprint arXiv:2402.03710*.

Peter Knees and Markus Schedl. 2013. A survey of music similarity and recommendation from music context data. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 10(1):1–21.

Younghun Lee, Dan Goldwasser, and Laura Schwab Reese. 2024. Towards understanding counseling conversations: Domain knowledge and large language models. *arXiv preprint arXiv:2402.14200*.

Andreas C Lehmann, John A Sloboda, and Robert H Woody. 2007. *Psychology for musicians: Understanding and acquiring the skills*. Oxford University Press.

Jiajia Li, Lu Yang, Mingni Tang, Cong Chen, Zuchao Li, Ping Wang, and Hai Zhao. 2024. The music maestro or the musically challenged, a massive music evaluation benchmark for large language models. *arXiv preprint arXiv:2406.15885*.

Jinhua Liang, Huan Zhang, Haohe Liu, Yin Cao, Qiuqiang Kong, Xubo Liu, Wenwu Wang, Mark D Plumbley, Huy Phan, and Emmanouil Benetos. 2024. Wavcraft: Audio editing and generation with large language models. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Liwei Lin, Gus Xia, Junyan Jiang, and Yixiao Zhang. 2023. Content-based controls for music large language modeling. *arXiv preprint arXiv:2310.17162*.

Liwei Lin, Gus Xia, Yixiao Zhang, and Junyan Jiang. 2024. Arrange, inpaint, and refine: Steerable long-term music audio generation and editing via content-based controls. *arXiv preprint arXiv:2402.09508*.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2024. Music understanding llama: Advancing text-to-music generation with question answering and captioning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 286–290. IEEE.

Xubo Liu, Zhongkai Zhu, Haohe Liu, Yi Yuan, Meng Cui, Qiushi Huang, Jinhua Liang, Yin Cao, Qiuqiang Kong, Mark D Plumbley, et al. 2023. Wavjourney: Compositional audio creation with large language models. *arXiv preprint arXiv:2307.14335*.

Yin-Jyun Luo, Kat Agres, and Dorien Herremans. 2019. Learning disentangled representations of timbre and pitch for musical instrument sounds using gaussian mixture variational autoencoders. *arXiv preprint arXiv:1906.08152*.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. 2024. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. *arXiv preprint arXiv:2404.09956*.

Pranay Manocha, Adam Finkelstein, Richard Zhang, Nicholas J Bryan, Gautham J Mysore, and Zeyu Jin. 2020. A differentiable perceptual audio metric learned from just noticeable differences. *arXiv preprint arXiv:2001.04460*.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2024. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2024. Mustango: Toward controllable text-to-music generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8286–8309.

Leonard B Meyer. 2008. *Emotion and meaning in music*. University of chicago Press.

Michele Newman, Lidia Morris, and Jin Ha Lee. 2023. Human-ai music creation: Understanding the perceptions and experiences of music creators for ethical and productive collaboration. In *ISMIR*, pages 80–88.

Javier Nistal, Marco Pasini, Cyran Aouameur, Maarten Grachten, and Stefan Lattner. 2024. Diff-a-riff: Musical accompaniment co-creation via latent diffusion models. *arXiv preprint arXiv:2406.08384*.

Emilian Postolache, Giorgio Mariani, Luca Cosmo, Emmanouil Benetos, and Emanuele Rodolà. 2024. Generalized multi-source inference for text conditioned music diffusion models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6980–6984. IEEE.

Simeon Rau, Frank Heyen, Stefan Wagner, and Michael Sedlmair. 2022. Vi-sualization for ai-assisted composing. In *Proc. of the 23rd Int. Society for Music Information Retrieval Conf.*

Francesca Ronchini, Luca Comanducci, Gabriele Perego, and Fabio Antonacci. 2024. Paguri: a user experience study of creative interaction with text-to-music models. *arXiv preprint arXiv:2407.04333*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2024. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36.

John A Sloboda. 1986. *The musical mind: The cognitive psychology of music*. Oxford University Press.

Or Tal, Alon Ziv, Itai Gat, Felix Kreuk, and Yossi Adi. 2024. Joint audio and symbolic conditioning for temporally controlled text-to-music generation. *arXiv preprint arXiv:2406.10970*.

Ziyu Wang, Dingsu Wang, Yixiao Zhang, and Gus Xia. 2020. Learning interpretable representation for controllable polyphonic music generation. *arXiv preprint arXiv:2008.07122*.

Alec Wright and Vesa Välimäki. 2020. Perceptual loss function for neural modeling of audio systems. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 251–255. IEEE.

Ho-Hsiang Wu, Oriol Nieto, Juan Pablo Bello, and Justin Salamon. 2023. Audio-text models do not yet leverage natural language. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J Bryan. 2024. Music controlnet: Multiple time-varying controls for music generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2692–2703.

Hiromu Yakura and Masataka Goto. 2023. Iteratta: An interface for exploring both text prompts and audio priors in generating music with text-to-audio models. In *Ismir 2023 Hybrid Conference*.

Yi-Hsuan Yang and Homer H Chen. 2012. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):1–30.

Yao Yao, Peike Li, Boyu Chen, and Alex Wang. 2023. Jen-1 composer: A unified framework for high-fidelity multi-track music generation. *arXiv preprint arXiv:2310.19180*.

Dingyao Yu, Kaitao Song, Peiling Lu, Tianyu He, Xu Tan, Wei Ye, Shikun Zhang, and Jiang Bian. 2023. Musicagent: An ai agent for music understanding and generation with large language models. *arXiv preprint arXiv:2310.11954*.

Ruibin Yuan, Hanfeng Lin, Yi Wang, Zeyue Tian, Shangda Wu, Tianhao Shen, Ge Zhang, Yuhang Wu, Cong Liu, Ziya Zhou, et al. 2024a. Chatmusician: Understanding and generating music intrinsically with llm. *arXiv preprint arXiv:2402.16153*.

Yi Yuan, Zhuo Chen, Xubo Liu, Haohe Liu, Xuenan Xu, Dongya Jia, Yuanzhe Chen, Mark D Plumbley, and Wenwu Wang. 2024b. T-clap: Temporal-enhanced contrastive language-audio pretraining. *arXiv preprint arXiv:2404.17806*.

Yixiao Zhang, Yukara Ikemiya, Woosung Choi, Naoki Murata, Marco A Martínez-Ramírez, Liwei Lin, Gus Xia, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. 2024a. Instruct-musicgen: Unlocking text-to-music editing for music language models via instruction tuning. *arXiv preprint arXiv:2405.18386*.

Yixiao Zhang, Yukara Ikemiya, Gus Xia, Naoki Murata, Marco Martínez, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. 2024b. Musicmagus: Zero-shot text-to-music editing via diffusion models. *arXiv preprint arXiv:2402.06178*.

Yixiao Zhang, Akira Maezawa, Gus Xia, Kazuhiko Yamamoto, and Simon Dixon. 2023. Loop copilot: Conducting ai ensembles for music generation and iterative editing. *arXiv preprint arXiv:2310.12404*.

Yixiao Zhang, Gus Xia, Mark Levy, and Simon Dixon. 2021. Cosmic: A conversational interface for human-ai music co-creation. In *NIME 2021*. PubPub.