Medley

ECE1786 Final Report

December 12, 2023

Ze Wang 1004065502
Zhaoyu Yan 1006521621

Word Count: 1742
Word Penalty: 0%

Introduction

Music is an important part of our lives, but finding the exact songs you want is challenging. Traditional music recommendation systems often rely on users' previous listening history or search based on genres or artists, which might not always align with their current preferences or mood. This project aims to develop a song recommendation system that understands complex user requests and suggests the top matching songs. Users can input specific music preferences, such as "recommend pop songs from artists similar to Justin Bieber, in English, with a happy mood," and receive a list of recommended songs.

Illustration

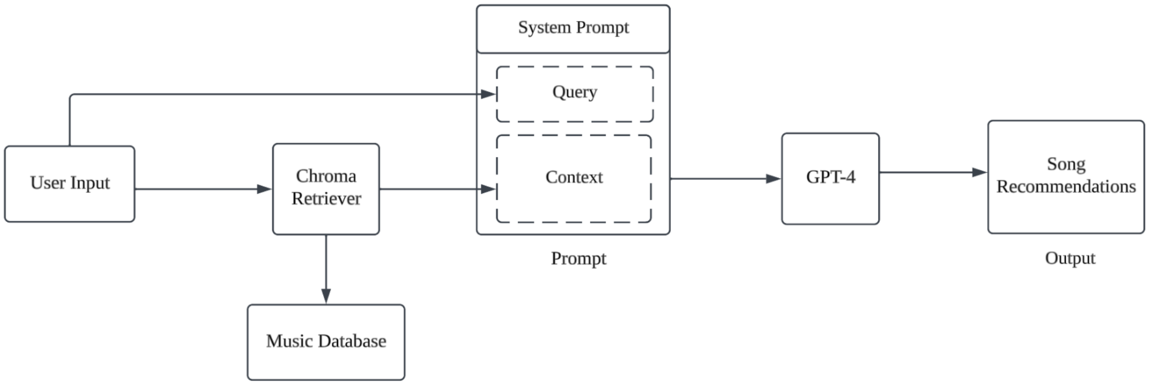


Figure 1: System Structure

A "Rock Classic" example:

User Input:
Music genres: Classic Rock, Hard Rock, Blues Rock
Artists: The Who, Boston, Pink Floyd
Languages: English
Contextual uses: Nostalgic anthems, road trips, rock concerts
Mood: Energetic, nostalgic, rebellious, introspective, passionate

Recommendation:

1. "More Than a Feeling" by Boston
2. "Sweet Home Alabama" by Lynyrd Skynyrd
3. "Baba O'Riley" by The Who
4. "Take It Easy" by Eagles
5. "Ramblin' Man" by The Allman Brothers Band
6. "American Girl" by Tom Petty and the Heartbreakers
7. "The Weight" by The Band
8. "Simple Man" by Lynyrd Skynyrd

9. "Peace of Mind" by Boston
10. "Behind Blue Eyes" by The Who

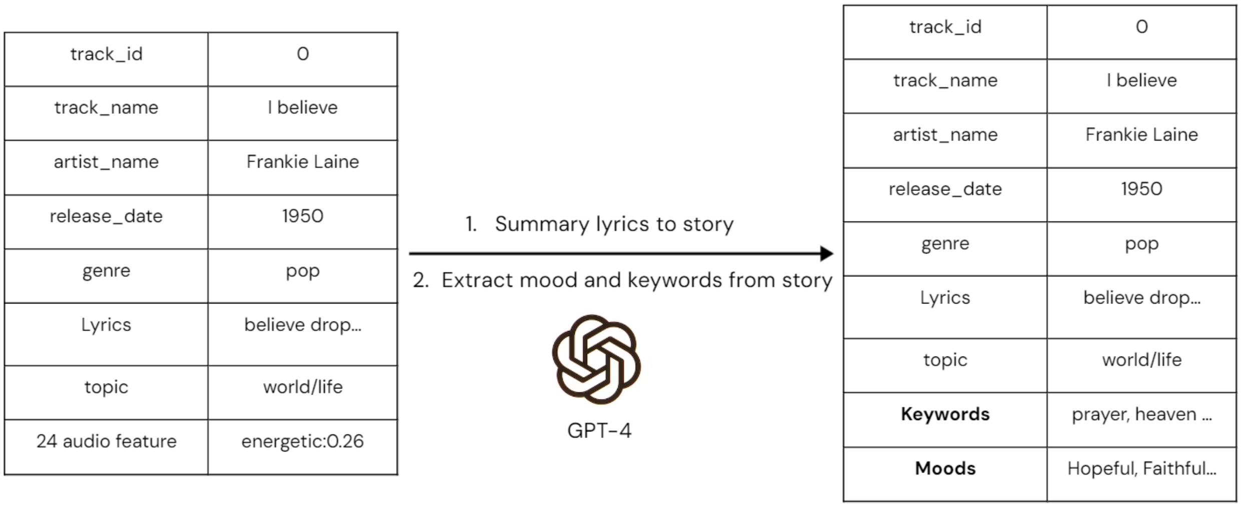
Background & Related Work

PALR: Personalization Aware LLMs for Recommendation [2]: The core approach of this paper is to enhance recommendation systems by integrating user history with Large Language Models (LLMs). The PALR framework operates in three stages: generating a user profile by summarizing a user's preferences using an LLM, filtering candidate items or services based on this profile, and finally, using the LLM to generate personalized recommendations from these candidates. This method leverages the advanced language processing capabilities of LLMs to provide more accurate and tailored recommendations, demonstrating its effectiveness through various sequential recommendation tasks.

Recommender Systems in the Era of Large Language Models (LLMs) [1]: This paper explores the integration of LLMs, like ChatGPT and GPT-4, into recommender systems. The approach involves a comprehensive review of LLM-empowered recommender systems, focusing on how LLMs can be used as feature encoders for learning user and item representations. It also discusses advanced techniques for enhancing recommender systems and addresses the limitations of existing systems. The paper highlights the potential of LLMs in improving the personalization and accuracy of recommendations, providing valuable insights for the enhancement of recommender systems using LLMs.

Data and Data Processing

In this project, two main datasets have been collected. The first dataset, known as the 'Music Dataset: 1950 to 2019' [4], includes approximately 27,000 music tracks spanning from 1950 to 2019. Although each track in this dataset is characterized by 35 different features, only 7 were employed in this study. The project team has applied GPT-4 for transforming the lyrics of these tracks into concise narratives, aiming to reveal the underlying contexts and emotions. This approach enabled the extraction of keywords and moods from each track. For example, the track "I Believe" by Frankie Laine is identified with keywords such as 'prayer' and 'heaven', and moods like 'hopeful' and 'faithful'. This processed collection of data is henceforth referred to as the ‘Music Dataset'.



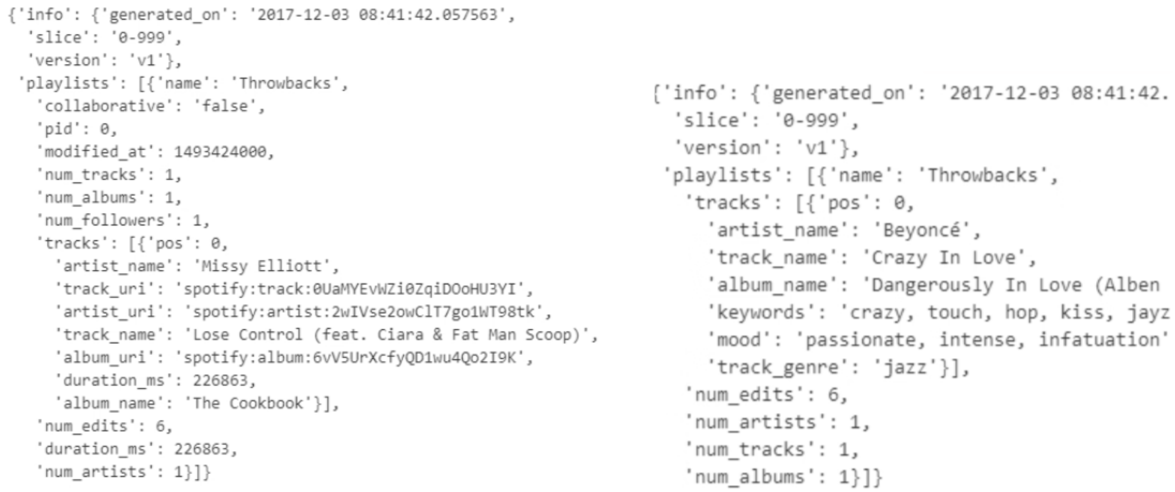
track_id	0
track_name	I believe
artist_name	Frankie Laine
release_date	1950
genre	pop
Lyrics	believe drop...
topic	world/life
24 audio feature	energetic:0.26

1. Summary lyrics to story
2. Extract mood and keywords from story
GPT-4

track_id	0
track_name	I believe
artist_name	Frankie Laine
release_date	1950
genre	pop
Lyrics	believe drop...
topic	world/life
Keywords	prayer, heaven ...
Moods	Hopeful, Faithful...

Figure 2: Example of a track in the Music Dataset before and after processed

The second pivotal dataset in this research is the 'Spotify Million Playlist Dataset' [3], which is used to simulate user inputs. Non-essential features such as playlist duration have been filtered out, and tracks not aligning with those in the Music Dataset, as well as playlists containing fewer than 10 tracks, have been excluded. This curated dataset, known as the 'Playlist Dataset', does not include every track from the original dataset. Despite this limitation, the dataset offers significant insights into the capabilities of language models in the realm of music recommendation systems.



```
{'info': {'generated_on': '2017-12-03 08:41:42.057563',
  'slice': '0-999',
  'version': 'v1'},
 'playlists': [{'name': 'Throwbacks',
   'collaborative': 'false',
   'pid': 0,
   'modified_at': 1493424000,
   'num_tracks': 1,
   'num_albums': 1,
   'num_followers': 1,
   'tracks': [{'pos': 0,
     'artist_name': 'Missy Elliott',
     'track_uri': 'spotify:track:0UaMYEvWZi0ZqiDOoHU3YI',
     'artist_uri': 'spotify:artist:2wIVse2owClT7go1WT98tk',
     'track_name': 'Lose Control (feat. Ciara & Fat Man Scoop)',
     'album_uri': 'spotify:album:6vV5UrXcfyQD1wu4Qo2I9K',
     'duration_ms': 226863,
     'album_name': 'The Cookbook'}],
   'num_edits': 6,
   'duration_ms': 226863,
   'num_artists': 1}]}
```

```
{'info': {'generated_on': '2017-12-03 08:41:42.
  'slice': '0-999',
  'version': 'v1'},
 'playlists': [{'name': 'Throwbacks',
   'tracks': [{'pos': 0,
     'artist_name': 'Beyoncé',
     'track_name': 'Crazy In Love',
     'album_name': 'Dangerously In Love (Alben
     'keywords': 'crazy, touch, hop, kiss, jayz
     'mood': 'passionate, intense, infatuation'
     'track_genre': 'jazz'}],
   'num_edits': 6,
   'num_artists': 1,
   'num_tracks': 1,
   'num_albums': 1}]}
```

Figure 3: Example of a playlist in the Playlist Dataset before and after processed

Architecture and Software

The system, as illustrated in Figure 1, integrates Retrieval-Augmented Generation using the LangChain framework. For retrieval, the OpenAI embedding was leveraged to embed the Music Dataset. This vectorized data is then stored in the Chroma database. When a user inputs a query, the Chroma retriever scans this database to identify and fetch the 15 most relevant songs, based on their similarity to the user's query.

The generation component utilizes the OpenAI GPT-4 chat model. It processes two types of prompts: the user prompt and the system prompt. The user prompt is a fusion of the original query and the context retrieved from the Chroma database. The system prompt, which is consistent across all queries, first explains what the given input contains and then directs GPT-4 to function as a recommendation system, recommending 10 published, distinct songs that align with the user's request and the retrieved context. This ensures no hallucination or duplication in the output.

Instead of directly recommending songs from the retrieved list, GPT-4 was prompted to use these songs as examples for generating its own recommendations. This is because we found that GPT-4 has extensive knowledge of existing songs and when we compared the performances of using RAG alone, GPT-4 without RAG, and GPT-4 with RAG, it was evident that GPT-4 with RAG outperformed the other setups. It might be that the integration of RAG aids GPT-4 in effectively understanding ambiguous user requests.

Greedy decoding is employed, setting the temperature to 0, to ensure that the model's outputs are precise and definitive. Greedy decoding selects the most likely next word at each step, reducing the chance of irrelevant or repetitive recommendations.

Comparison

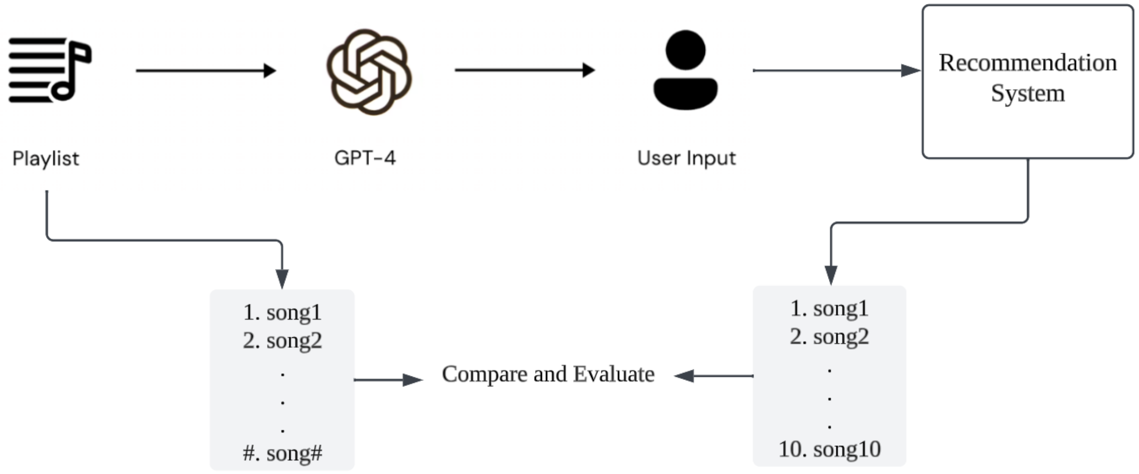


Figure 4: Flowchart of synthesizing user input for testing

To test the system performance in the absence of an extensive dataset of real user requests, the GPT-4 model is utilized to analyze playlists in the Playlist Dataset. GPT-4 is prompted to summarize and extract key features such as genres, artists, languages, keywords, and mood. The model then uses this extracted information to generate synthetic user requests. These requests are designed to mimic what a user, who created a given playlist, might ask the system. Around 20 generated outputs are examined to ensure the results are appropriate. This approach also allows us to compare the system's outputs with the original playlists to assess accuracy and relevance.

In this project, performance evaluation is conducted on both quantitative and qualitative measures. The quantitative analysis used the Hit-Rate at 10 (HR@10) metric, which measures the percentage of times a relevant item appears in the top 10 recommendations. This provides insight into the accuracy of our recommendation system. For qualitative assessments, we closely examined the suitability of the recommendations in relation to the unique context, mood, and character of the original playlist. This involved evaluating how well the recommendations matched the emotional tone and thematic elements of the playlist.

Quantitative Results

HR@10, a standard metric for assessing recommendation systems, evaluates whether at least one of the top 10 recommended items is an exact match. In our project, a 'hit' is defined as an exact match of the track, and a 'hit recommendation' refers to a recommendation that contains at least one hit. If so, it scores a 'hit' and returns 1; otherwise, it scores 0. In our project, we applied this simplified formula.

$$\text{HR@10} = \frac{\#\text{ of hit track}}{\#\text{ of playlist}} = \frac{\text{128 hits tracks}}{\text{200 random playlist}} = 64\%$$

$$\text{Average }\#\text{ hit track} = \frac{\text{total }\#\text{ of hit tracks}}{\text{total }\#\text{ of playlist}} = \frac{\text{249 hit track}}{\text{200 Random playlist}} = 1.245$$

Our model achieved a 64% success rate on the HR@10 metric, averaging 1.245 hits per playlist. This was confirmed through tests on 200 randomly chosen playlists.

Qualitative Results

For qualitative evaluation in this project, the focus was on determining if the recommended genre or artist aligns with the context and unique characteristics of the playlist. The results indicate that even if some recommendations are not direct hits, they can still be considered suitable matches. This demonstrates the system's effectiveness in aligning recommendations with personalized user requirements, highlighting its ability to understand and cater to specific user preferences and contexts.

For instance, the best recommendation, 8 hits out of 10, the 'Rock Classics' playlist featuring classic rock from the 60s and 70s, a track like 'The Weight' by The Band is deemed suitable match for its thematic alignment, while 'American Girl' is seen as unsuitable due to thematic dissonance.

Track Name	Artists	Release year	Genre	Topic	Suitable
The Weight	The Band	1968 √	rock classic √	responsibility√	√
American Girl	Tom Petty	1786 √	rock classic √	love ×	×

Table 1: Qualitative analysis on two not-hit songs in a randomly selected best recommendation

In the final analysis of random zero-hit recommendations, the 'Pregame' playlist, predominantly featuring high-tempo rap tracks for pre-game energizing, was examined. In this context, 'Stronger' by Kanye West was a fitting recommendation, aligning with the playlist's high-energy and motivational theme, despite not being a direct hit. Conversely, a track like 'Turn Down For What' by DJ Snake and Lil Jon was considered unsuitable due to a mismatch in genre, demonstrating the nuanced understanding of the recommendation system in aligning with the specific mood and style of a playlist.

For a comprehensive analysis of the full set of recommendations for the 'Pregame' playlist , please refer to the appendix. 5 recommendation is considered suitable match dispute not a direct hit.

Track Name	Artists	Tempo	Genre	Topic	Suitable
Stronger	Kanye West	High √	Rap √	self-empowerment √	√
Turn Down for What	DJ Snake & Lil Jon 0	High √	Electronic Dance Music ×	refusal to stop √	×

Table 2: Qualitative analysis on two not-hit songs in a randomly selected worst recommendation

Discussion and Learnings

The results exceeded our expectations. While the HR@10 was only 0.64, indicating a moderate level of accuracy, the system demonstrated an acceptable level of diversity in its recommendations. The qualitative analysis revealed that many recommendations not present in the original playlists were still highly relevant and potentially appealing to users.

During the project, there were some surprising findings. One is that GPT-4 has a notable understanding of existing songs and their artists, likely a result of its extensive training data. The model was particularly good at making recommendations for topic-specific requests, such as classic rock and Disney songs, even without the use of RAG. This might be attributed to the more predictable nature of less diversified songs. Additionally, its recommendations often align with the asked genres and artists, suggesting its effective use of user request features.

However, several limitations were observed. GPT-4 struggled with user requests that feature a mix of music styles, possibly due to the lack of clear thematic cues. The presence of unrelated tracks in the recommendations highlighted a limitation in the model's contextual understanding. Furthermore, it showed limited grasp of the emotional and narrative context behind songs, which could be attributed to the absence of this specific information in its training data.

Future improvements would focus on incorporating additional factors such as collaborations among artists and the popularity of songs into the RAG, as these elements are crucial in reflecting people's musical preferences. Also, integrating a user feedback loop would enable more personalized and refined recommendations.

Individual Contributions

Ze’s contributions:

- Collected the Music Dataset
- Processed and cleaned the datasets
- Refined the system prompt for generating song recommendations (50%)
- Wrote code for running and evaluating the system
- Conducted quantitative evaluation of the system on the testing data
- Participated in qualitative analysis of the generated results (50%)

Zhaoyu’s contributions:

- Collected the Playlist Dataset
- Was responsible for generating synthetic user requests
- Refined the system prompt for generating song recommendations (50%)
- Wrote code for using GPT-4 through the OpenAI API
- Wrote the RAG implementation
- Participated in qualitative analysis of the generated results (50%)

Reference

[1] W. Fan et al., “Recommender Systems in the Era of Large Language Models (LLMs),” 2023, doi: 10.48550/arxiv.2307.02046.

[2] F. Yang, Z. Chen, Z. Jiang, E. Cho, X. Huang, and Y. Lu, “PALR: Personalization Aware LLMs for Recommendation,” 2023, doi: 10.48550/arxiv.2305.07622.

[3] C.-W. Chen, P. Lamere, M. Schedl, and H. Zamani, "Recsys challenge 2018: automatic music playlist continuation," in RecSys '18: Proceedings of the 12th ACM Conference on Recommender Systems, September 2018, pp. 527-528, doi: 10.1145/3240323.3240342

[4] L. Moura, E. Fontelles, V. Sampaio, and M. França, "Music Dataset: Lyrics and Metadata from 1950 to 2019," Mendeley Data, vol. V3, 2020. [Online]. Available: https://doi.org/10.17632/3t9vbwxgr5.3

Appendix

Lyric prompt:
“Summarize the provided song lyrics by focusing on the main ideas and essential information. Analyze the mood and context of the song, and extract the most significant simple keywords. Compare these keywords to determine the top five that best represent the lyrics.
ONLY OUTPUT SONG NAME, KEYWORDS and MOOD. Output format: Name: Keywords: 1, 2., 3., 4., 5. Mood: 1, 2., 3."

Playlist prompt:
“Analyze a user's playlist to deduce musical preferences and mood. Given the playlist name, and a list of tracks with corresponding artist names:
1. Determine the user's top three favourite music genres based on the tracks' styles.
2. Identify the user's top three favourite artists from the playlist.
3. Ascertain the user's preferred language of music, indicating regional music preferences if any.
4. Infer the contextual use of the playlist by identifying any patterns that suggest particular events, locations, or themes.
5. Assess the overall mood of the playlist. Categorize and summarize the mood based on the tone and tempo of the songs into 5 words.
Compile the findings into a profile summary that shows the user's musical tastes.”

System prompt:
“As a music recommendation system, provide tailored song recommendations based on detailed user inputs. Given a request containing:
- The user's specific music preferences, such as genres and artists.
- The user's language preferences for music.

- The context or theme preferences for songs
- The user's mood preferences for songs.

And a list of examples of relevant songs to the request.
Recommend 10 unique, published songs that align with the user's musical preferences.
Present the recommendations in a list format, with each entry following the '[song_name] by [artist_name]' structure.”

‘Pregame’ playlist complete analyze:

Track Name	Artists	Tempo	Genre	Topic	Suitable
Stronger	Kanye West	High √	Rap √	self-empowerment √	√
Black and Yellow	Wiz Khalifa	High √	Hip-Hop/Rap √	pride, success √	√
SICKO MODE	Travis Scott	High √	Hip-Hop/Trap √	Success √	√
HUMBLE	Kendirck Lamar	High √	Hip-Hop/Rap √	Self-awareness √	√
No Problem	Chance the Rapper ft. Lil Wayne & 2 Chainz	High √	Hip-Hop/Rap √	independence and defiance √	√
Turn Down for What	DJ Snake & Lil Jon 0	High √	Electronic Dance Music ×	refusal to stop √	×
Uptown Funk	Mark Ronson ft. Bruno Mars	moderately fast ×	Funk, Pop ×	Confident √	×
POWER	Kanye West	High √	Hip-Hop/Trap √	burdens of success ×	×
Lose Control	Missy Elliott ft. Ciara & Fat Man Scoop	High √	Hip-Hop/Rap	Enjoying party ×	×
I Gotta Feeling	The Black Eyed Peas	moderate ×	Pop	enjoyment ×	×

Permissions

Team Member	Post Video?	Post Final Report?	Post Source Code?
Ze Wang	Yes	Yes	Yes
Zhaoyu Yan	Yes	Yes	Yes