

Taylor’s Tunes

Mood-Driven Music Recommendations with Large Language Models

Group 4

Duong Le: duong.h.le@aalto.fi
Hieu Pham: hieu.pham@aalto.fi
Tommaso Canova: tommaso.canova@studenti.unitn.it

February 2024

Abstract

The project aims to create an end-to-end product that is capable of recommending songs based on the mood of its users. We plan to use open source Large Language Models such as *Mixtral-8x7b*, *LLaMA2-70b-chat*, *Gemma-7b-it* and to leverage Retrieval Augmented Generation (RAG) using lyrics from the artist Taylor Swift’s songs to be able to understand human emotions and recommend pieces of music accordingly. We are planning to store the lyrics using a vector database such as *Qdrant* and to orchestrate the RAG pipeline using *Langchain*.

The user would start a conversation with Taylor’s Tunes. Throughout the conversation, the user describes how they are feeling and Taylor’s Tunes would capture these emotions based on the context or explicitly from the user’s text. At the end of the conversation, Taylor’s Tunes outputs a Taylor Swift song that matches the best with the emotion captured during the session.

Literature Review

Taylor Swift dataset

A tailored dataset on Taylor Swift songs has been proposed in *Mansfield* and *Seligman* work [1]. The authors of the paper put a specific focus on highlighting the level of happiness/optimism and strength commitment to a relationship based on the lyrics and chordal tones of each song (180 in total). Analysing the songs a custom marker indicated as *MIQ* (“Male in Question”) has been proposed to address the song toward Taylor Swift speaking directly, or in an indirect way by expressing her feelings. The happiness score is composed of four criteria, most of them can reach a score between -3 and +3. The first criterion

encapsulates Taylor Swift self feelings in the song, the second is referred to as "Glass Half Full" and measures the song author's outlook on life. The third one is split into two halves: the negative one refers to negative emotions such as depression, anger, denial while the positive half embodies positive emotions. The fourth one measures the tempo and musical feel of the song based on its key and BPM, while a final ± 3 score was reserved for songs with either positive or negative "hyperbolic line" that lead to a sudden decrease or increase of emotional intensity.

With a similar structure, the relationship Criteria is proposed. "Seriousness of Topics Discussed" takes in consideration how committed the author and the MIQ were to each other based on what topics they discussed together. The second criterion takes into consideration the future prospects in the relationship. the third criterion quantifies how the MIQ feels about Taylor Swift, while the final one is how much time Taylor and the MIQ seem to spend together throughout the course of the song.

A scoring example for one of the happiness criteria is given in Table 1

Table 1: Glass Half Full: one of four happiness criteria

Score	Description	Example
-3	All imagery is depressing	"And then the cold came, the dark days when fear crept into my mind" (<i>Back To December</i>)
-2	Nearly all depressing imagery	"How you laugh when you lie / You said the gun was mine / Isn't cool, no I don't like you / But I got smarter, I got harder" (<i>Look What You Made Me Do</i>)
-1	Majority depressing imagery	"Stealing hearts and running off and never saying sorry / But if I'm a thief then / He can join the heist" (<i>...Ready For It?</i>)
0	Equal amounts of happy and sad imagery	"Rain came pouring down when I was drowning / That's when I could finally breathe" (<i>Clean</i>)
1	Majority positive imagery	"We're happy, free, confused, and lonely in the best way / It's miserable and magical" (<i>22</i>)
2	Nearly all positive imagery	"This love is difficult, but it's real / Don't be afraid, we'll make it out of this mess / It's a love story, baby just say yes" (<i>Love Story</i>)
3	All imagery is positive	"And all I feel in my stomach is butterflies / The beautiful kind, making up for lost time" (<i>Everything Has Changed</i>)

Vector Database and Retrieval Augmented Generation

To process the lyrics data, we need to find a way to represent said data. There are many ways to achieve this such as One-hot Encoding, Bag-of-Words, TF-IDF, etc... [2], but the end results would always be numerical vectors. Thus,

it makes sense that we utilized a database system that can effectively store and retrieve such vectors. That is where Vector Database Management System (VDBMS or just Vector Database for short) [3] comes in. Essentially, VDBMS is a functional software that focus on effectively manage high dimensional vectors. VDBMS is optimized to search for similar vectors to the queried vectors rather than to look for the perfect match. This can be done effectively through indexing the data in the database. Current challenges with VDBMS include speed-accuracy trade-off, growing dimensionality and sparsity, and general maturity of the system. In our project, VDBMS can be used to store the lyrics and metadata of the songs to use in retrieval augmented generation (RAG).

RAG is a method used applied to text generative models with the aim of providing said model a non-parametric memory that the model can retrieve data from [4]. Traditional text generative models are able to gain knowledge from data [5]. However, they struggle to provide context for said knowledge and can suffer from hallucinations, where the model presents misinformation as truth [6]. Through the use of RAG, this problem can be mitigated by giving the model a method to revise and retrieve correct information [4]. An illustration of how RAG works can be seen in Figure 1. The model in the illustration can identify and retrieve relevant information from its memory. These actions can be done through the use of a Vector Database as explained above. Compared to other fine-tuning approaches, the former is analogous to giving the model a reference list to help it find the relevant information, while the latter is similar to teaching the model additional knowledge. Both have their own advantages, but for our application, RAG is a better choice. There are three types of RAG: *naive*, *advanced*, and *modular* RAG [7], as illustrated in Figure 2. They are characterized by the use of pre and post retrieval processes, as well as additional modules in the case of modular RAG. In our project, we used naive RAG due to its simplicity in implementation, but future iterations of the projects are encouraged to try the other types.

In the context of our project, RAG can be used to help Taylor’s Tunes provide reasoning for the chosen song. In addition to the selected song, Taylor’s Tunes can also present parts of the lyrics that are directly related to the detected emotions. These song parts can help the users evaluate the results of the LLM and request another song if unsatisfied. Currently, the metadata stored are the name, the lyrics, and the album of the songs, in which the name is used as the index. We stored the full lyrics, but in future steps of the project, we will switch to using the verses of the song. This is because strings are usually chunked before embedding and storing, and the verses of the songs are a natural way to split the song into parts.

LLMs Evaluation

Taylor’s Tunes is an attempt to capture the emotions and thoughts of its user through the means of communication and use its understanding to interact with the user. Thus, to assess its performance means we need to be able to quantify its Emotional Intelligence (EI), which is necessary for effective communication

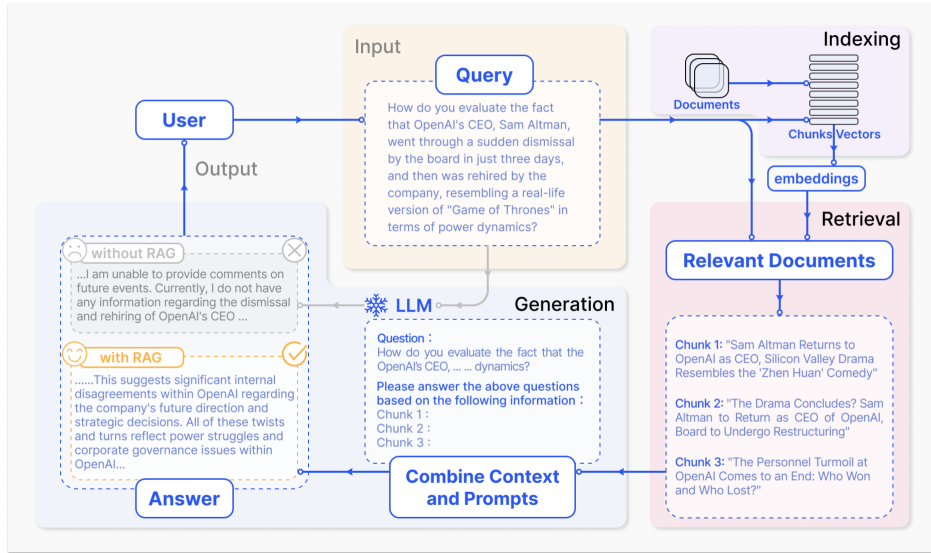


Figure 1: An illustration of how a text generative model uses RAG to answer a question posed by the user. Figure taken from [7]

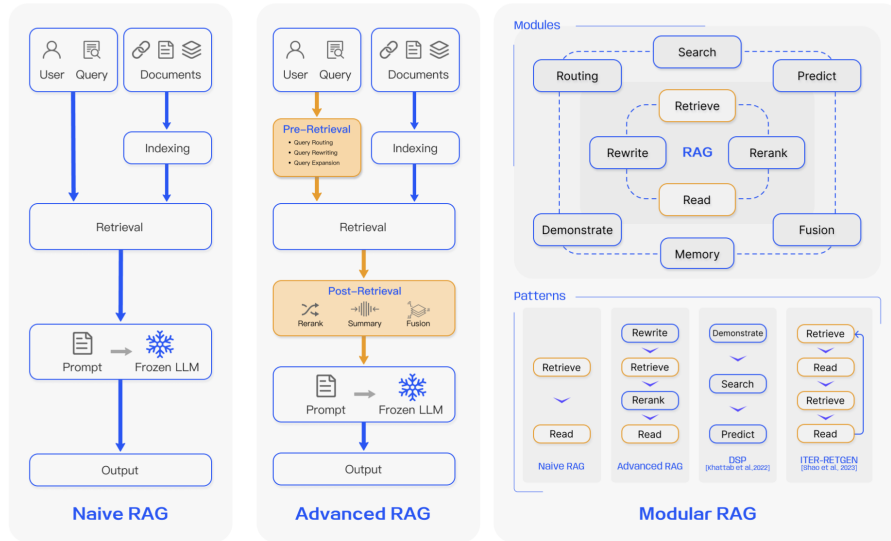


Figure 2: Comparison between the three paradigms of RAG. Figure taken from [7]

and social interactions [8]. Traditionally, the Theory of Mind (ToM) test has been applied to measure the ability to understand and correctly depict another's

mental state. LLMs before 2022 mostly showed no ability of ToM, being on par with small children’s ability [8]. ToM is a questionnaire ranging from false belief which means people have different understandings that diverge from reality [9], to pragmatic reasoning [10]. Due to ToM’s heterogeneous nature, it has been considered to not meet the reliability and validity standards to act as a psychometric test and has not been included in standardized tests on EI [11]. Wang et al. [8] has proposed a standardized test for Emotional Understanding (EU), which is a core component of EI, to be suitable for both humans and LLMs, termed the Situational Evaluation of Complex Emotional Understanding (SECEU). The test requires participants to evaluate complex emotions in realistic scenarios expressed by sentences. Specifically, the study conducted the test on both human volunteers and LLMs through 40 questions, each has 4 options of complex emotions. Participants were then asked to give the score for each emotion that sums up to 10. An example of such question and options can be seen in Figure 3. LLMs’ answers were compared with humans’ answers to see the similarities. In our study, due to the need to understand emotions through communicating with users, the original data from the SECEU study do not fit the purpose which is not a story but a conversation. Thus, we evaluated Taylor’s Tunes with a modified version of SECEU. Instead of using the same 40 questions, we have chosen data from Saif.M et al. [12], with engineered labels to be one-hot vectors of emotions: anger, fear, joy, sadness.

The SECEU test		The Standard Scores:			
Item 1 Story: The airplane model that Wang made fell from the sky one minute after take-off. When she inspected the model, she found a part that could possibly be improved. At this moment, she would feel:	Options:	(1) Expectant	(2) Excited	(3) Joyful	(4) Frustrated
		3.56	3.09	1.78	1.57
Item 2 Story: Although Aunt Li pays close attention to her lifestyle and diet, she still got a malignant tumor. The chances of curing the tumor now are quite slim, and Aunt Li can do nothing about it. She would feel:	Options:	(1) Desperate	(2) Fear	(3) Helpless	(4) Sad
		3.45	2.04	2.17	2.33
⋮		⋮			
		40 × 4 matrix			

Figure 3: Exemplars of the original SECEU test and the standard scores from the human responses

Project Plan

Firstly, we want to understand the main topics covered by the artist in her songs, for this reason we will perform a brief Explorative Data Analysis over the lyrics of the songs, trying to extract the most relevant words or themes for each album. Moreover, a brief analysis of the emotions of the songs will be carried on using the *NRC Emotion Lexicon* [13] which contains 27k terms, each one associated with an emotion among: fear, anger, anticipation, trust, surprise, positive, negative, sadness, disgust, joy.

In the second stage we will build a real-time chat-like application using an open-source Python library for building interactive web applications called *Streamlit*. To interact with the LLMs we will rely on GroqCloud API calls. This service provides three open source LLMs (*LLaMA2-70b chat*, *Mixtral-8x-7b* and *Gemma-7b-it*) and guarantees 15x faster inference speed compared to traditional cloud hardware thanks to their *Groq LPU™ inference Engine*, a cutting-edge hardware developed for these purposes. Subsequently, we will connect the Groq API with our application, and in the following step the vector database, using the orchestration tool *Langchain*, with which we can also handle the contextual chat memory.

To improve the Groq model, we will perform RAG using the data on the lyrics of the song, as well as the happiness score of said songs. In this part, we will combine the dataset of *Mansfield* and *Seligman* and the lyrics dataset. The lyrics in the lyrics dataset will be pre-splitted into verses. The lyrics and emotional data will then be vectorized and store in our Qdrant environment. This environment will be integrated into the Groq model to serve as its external memory for retrieving data.

Finally, we evaluate our model with two goals in mind: to measure its ability to understand users' emotions from conversations, and to recommend Taylor Swift songs accordingly to the emotions. To synthesize actual conversations with users, we use and engineer the dataset of personal tweets [12] with emotion labels so that we have a validation dataset. Each data point in the dataset will include a tweet that resembles a speech (for example: "now that I have my future planned out, I feel so much happier."), and a one-hot vector of 4 columns: anger, fear, joy, and sadness. We will then feed the model with the prompts including those statements, and compare its output to the prelabeled data. Moreover, to have a baseline for our model, we plan to have other LLMs such as Llama, Koala, etc. go through the same test and compare their performances.

Once we reach a specific confidence level for the model's understanding of emotions, we want to measure its ability to recommend suitable songs from Taylor Swift to users. To evaluate this process, we plan to let the model answer the questionnaire from the `taylorswift` package based on the prompt of users [1]. The questionnaire will consist of 6 questions, each corresponding with a quantifiable emotion (Feeling of self, Stage of relationship, Seriousness, Future prospects, Feelings of male, Togetherness). Then, with the answers in mind, we will benchmark the song recommended by our model and the one from the package against the human evaluation provided in the Taylor Swift dataset.

References

- [1] M. Mansfield and D. Seligman, “I knew you were trouble: Emotional trends in the repertoire of taylor swift,” *arXiv preprint arXiv:2103.16737*, 2021.
- [2] R. Egger, “Text representations and word embeddings: Vectorizing textual data,” in *Applied data science in tourism: Interdisciplinary approaches, methodologies, and applications*. Springer, 2022, pp. 335–361.
- [3] T. Taipalus, “Vector database management systems: Fundamental concepts, use-cases, and current challenges,” *Cognitive Systems Research*, p. 101216, 2024.
- [4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.
- [5] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel, “Language models as knowledge bases?” *arXiv preprint arXiv:1909.01066*, 2019.
- [6] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.
- [7] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, “Retrieval-augmented generation for large language models: A survey,” *arXiv preprint arXiv:2312.10997*, 2023.
- [8] X. Wang, X. Li, Z. Yin, Y. Wu, and L. Jia, “Emotional intelligence of large language models,” 2023.
- [9] S. Baron-Cohen, A. M. Leslie, and U. Frith, “Does the autistic child have a “theory of mind” ?” *Cognition*, vol. 21, no. 1, pp. 37–46, 1985. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0010027785900228>
- [10] D. Sperber and D. Wilson, “Pragmatics, modularity and mind-reading,” *Mind & Language*, vol. 17, no. 1-2, pp. 3–23, 2002. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0017.00186>
- [11] J. D. Mayer, D. R. Caruso, and P. Salovey, “The ability model of emotional intelligence: Principles and updates,” *Emotion Review*, vol. 8, no. 4, pp. 290–300, 2016. [Online]. Available: <https://doi.org/10.1177/1754073916639667>
- [12] S. M. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, “Semeval-2018 Task 1: Affect in tweets,” in *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA, 2018.

- [13] S. M. Mohammad and P. D. Turney, “Crowdsourcing a word-emotion association lexicon,” *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.