

## ABSTRACT

### Evaluating and Refining Music Recommendation Systems: A Combined Study of Algorithmic Techniques and User Feedback

Victor Martinez Gil

Director: Rivas, Pablo, Ph.D.

This two-sided study explores improving music recommendations through both computational modeling and user interaction. In the first phase, the TF-IDF (Term Frequency–Inverse Document Frequency) algorithm is applied to song lyrics, generating recommendations based on thematic word patterns. ChatGPT is used to evaluate these recommendations, revealing strengths in identifying lyrical themes but also limitations in emotional and contextual understanding. To address these gaps, the second phase introduces an interactive survey where users engage with ChatGPT over ten rounds, providing binary feedback (like/dislike) to refine recommendations. Results show growing user satisfaction, though participants noted issues like repetitive suggestions and limited genre diversity. While most found ChatGPT’s final summaries accurate, they also expressed a need for richer feedback and better emotional context.

APPROVED BY DIRECTOR OF HONORS THESIS:

---

Dr. Pablo Rivas, School of Engineering and Computer Science

APPROVED BY THE HONORS PROGRAM:

---

Dr. Elizabeth Corey, Director

DATE: \_\_\_\_\_

EVALUATING AND REFINING MUSIC RECOMMENDATION SYSTEMS:  
A COMBINED STUDY OF ALGORITHMIC TECHNIQUES  
AND USER FEEDBACK

A Thesis Submitted to the Faculty of  
Baylor University  
In Partial Fulfillment of the Requirements for the  
Honors Program

By  
Víctor Martínez Gil

Waco, Texas

May 2025

## TABLE OF CONTENTS

Preface . . . . .	iii
Acknowledgments . . . . .	v
Chapter One: Background . . . . .	1
Chapter Two: Designing Lyric-Based Music Suggestions . . . . .	7
Chapter Three: Evaluating AI's Capabilities as a Music Recommender . . . . .	19
Chapter Four: Results and Reflections from User Interaction . . . . .	24
Chapter Five: Conclusion . . . . .	29
Bibliography . . . . .	33

## PREFACE

In an age where digital personalization is transforming the way we experience media, music recommendation systems have become essential tools in shaping how listeners discover new content. The aim of this thesis is to evaluate and refine such systems by combining computational models with direct user interaction, offering a holistic approach to improving personalization in music streaming platforms. This research is structured in two parts. Part 1 explores a computational framework using the TF-IDF (Term Frequency-Inverse Document Frequency) algorithm applied to lyrical content. It examines how analyzing lyrics, specifically thematic word patterns, can reveal personal music preferences and inform algorithmic recommendations. The system developed for this phase utilizes cosine similarity to identify songs with comparable lyrical themes and integrates ChatGPT as an evaluative tool. ChatGPT's ability to detect thematic coherence and critique recommendation quality has helped pinpoint both the strengths and limitations of a lyrics-based model. While the model proved effective in identifying repeated themes, it struggled to incorporate contextual and emotional nuance. By focusing exclusively on the words and phrases that are commonly repeated in a user's preferred tracks, this study aims to uncover deeper insights into personal music preferences in order to improve the accuracy of song recommendations (Goodfellow 2018).

Recognizing these limitations, Part 2 of this thesis introduces another layer of analysis through an empirical, user-centered approach. This section builds on the foundation laid in Part 1 by incorporating an interactive recommendation system in which participants engaged in ten iterative rounds of AI-driven music suggestions powered by ChatGPT. With binary feedback (like/dislike), users influence the AI's understanding of their preferences. This direct interaction allowed us to evaluate ChatGPT's adaptability and accuracy in generating personalized music profiles and revealed deeper insights into user satisfaction, diversity of recommendations, and emotional nuance. Together, these two aspects are deeply interconnected: Part 1 provides the algorithmic base and a controlled evaluation setting, while Part 2 contextualizes and validates those findings in real-world interactions. This dual approach of computational analysis coupled with human feedback offers a more complete understanding of the capabilities and challenges inherent in music recommendation systems. The goal is to contribute to the development of smarter, more responsive technologies that align with the nuanced ways people engage with and enjoy music.

## ACKNOWLEDGMENTS

I would like to express my sincere gratitude to the Honors Program for providing me with the opportunity and resources to conduct this research, which has been pivotal in my development as a professional in the field. I am also deeply thankful to my mentor, Dr. Rivas, for embarking on this project with me and offering steadfast guidance and support throughout the process. I would also like to thank the McNair Scholars Program, which provided me with the opportunity to surround myself with individuals who have played a significant role in shaping my interests and academic pursuits. I am especially grateful to Professor Mathew Aars and Dr. Lance L. Littlejohn for taking the time to read my paper and serve on my defense committee—their feedback and support have been truly appreciated. Lastly, I extend my heartfelt thanks to my family, which I love.

## CHAPTER ONE

### Background

Personalized recommendation systems have become essential tools in today's digital landscape, helping users navigate overwhelming amounts of content by tailoring suggestions to individual preferences. As the demand for personalization has grown, so has the range of techniques used to deliver it. In the context of music streaming, recommendation technologies aim to provide listeners with meaningful and relevant musical experiences, often shaping how users discover new songs and artists. This chapter introduces the foundational concepts and methods that underpin modern music recommendation systems. It begins by outlining traditional approaches that have formed the basis of many widely used platforms, and then shifts to more recent innovations driven by artificial intelligence. Together, these sections provide the necessary context for understanding the motivations and methodology behind the research presented in this thesis.

#### *Traditional Recommendation Techniques*

While new recommendation systems are constantly being developed, three main types are predominantly used by major companies: Collaborative Filtering, Content-Based Filtering, and Hybrid Recommendation Systems. Collaborative Filtering operates by analyzing user behavior and preferences, such as ratings or listening history, to suggest items that similar users have enjoyed. It leverages the idea that if two users share similar tastes, the recommendations suitable for one will likely be relevant to the other.



Content-based filtering, on the other hand, focuses on the characteristics of the items themselves. For music recommendation systems, this involves analyzing attributes such as lyrics, genre, tempo, or artist information. The system recommends songs with similar features to those the user has previously liked, which makes it highly personalized. This approach doesn't rely on other users' data, which can be advantageous when user-specific data is sparse or when the goal is to tailor recommendations closely to individual preferences. Hybrid recommendation systems combine the strengths of both collaborative and content-based filtering to create more robust and accurate recommendations (Lokesh 2019). For this research, we will be focusing on content-based filtering recommendation systems, since our objective is to explore how lyrical content and other song attributes can be analyzed to make personalized music recommendations.

The TF-IDF (Term Frequency-Inverse Document Frequency) algorithm is a widely used content-based filtering technique in text processing to evaluate the importance of a word in a document relative to a collection of documents, or corpus (Rajaraman 2014). It combines two metrics: Term Frequency (TF) and Inverse Document Frequency (IDF). Term Frequency measures how often a word appears in a specific document. It is calculated by dividing the number of times a word appears in the document by the total number of words in that document. This reflects the word's prominence within the particular text. Inverse Document Frequency on the other hand, assesses the importance of the word across the entire corpus. It is computed by taking the logarithm of the total number of documents divided by the number of documents containing the word. This adjustment helps to diminish the weight of common words that appear frequently across many documents, while highlighting words that are rare and

thus more significant within the specific document. The final TF-IDF score is obtained by multiplying these two metrics, providing a balanced measure of a word's relevance to the document while considering its general importance across the corpus (Goodfellow 2018).

The TF-IDF algorithm aids in vectorizing song lyrics by transforming textual data into numerical vectors that capture the importance of each word in relation to the entire corpus; Figure 1 depicts a general overview of the algorithm.

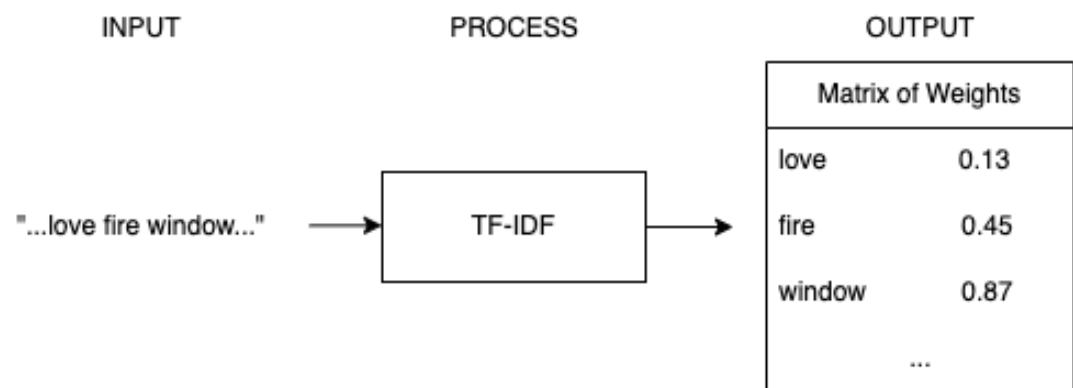


Figure 1. TF-IDF algorithm overview.

By calculating the TF-IDF score for each word, we generate a vector representation for each song where each dimension corresponds to a word in the vocabulary, and its value reflects the word's significance in that song as seen in Figure 1. This numerical representation allows us to systematically analyze and compare lyrics. To find the closest matching lyrics, we employ cosine similarity, which is a metric that measures the cosine of the angle between two vectors in a high-dimensional space. Cosine similarity calculates how aligned two vectors are, regardless of their magnitude, by evaluating the cosine of the angle between them. This method is effective in determining the similarity between two sets of lyrics as it focuses on the orientation of the vectors rather than their length. By computing the cosine similarity between the

vector of a user's preferred song and the vectors of other songs, we can identify and recommend the most similar lyrical content (Goodfellow 2018).

### *AI-Based Recommendation Techniques*

In recent years, the integration of artificial intelligence into recommendation systems has ushered in novel strategies for capturing user preferences with greater nuance and adaptability. Contemporary AI-based recommendation techniques contrast significantly with traditional models, which predominantly rely on historical data or static item attributes. Such traditional methods often render simplistic suggestions that lack the sophistication necessary to cater to the multifaceted nature of user preferences. In contrast, AI-driven systems leverage advanced machine learning algorithms and natural language processing technologies to discern complex patterns in user behavior, item characteristics, and contextual factors. For instance, these systems can account for variables such as the user's mood or time of day, thereby tailoring recommendations to fit varying user contexts (Schedl et al., 2018; Schedl, 2019).

The ability of AI systems to continuously learn and adapt is one of their paramount features, facilitating the refinement of recommendations based on real-time feedback and emerging trends. This aspect of AI recommendations is particularly noteworthy, as it enables systems to evolve alongside changing user preferences. Consequently, the algorithms that power these systems can ensure relevance in an ever-changing digital landscape (Schedl, 2019). This trend highlights a shift from static to dynamic recommendation models, illustrating how AI can enhance the user experience through ongoing engagement and feedback mechanisms.

One of the most transformative advancements in this domain has been the deployment of deep learning architectures, such as recurrent neural networks (RNNs) and transformers. These architectures revolutionize the modeling of sequential user behavior, allowing for the identification of long-term dependencies across historical interactions. For instance, RNNs are adept at capturing the temporal dynamics of user behaviors, which is crucial for applications such as music recommendations where the sequence of songs played significantly influences user satisfaction. This sequential modeling is essential for understanding intricate patterns in listening histories, thus leading to more personalized suggestions that resonate with users' evolving tastes (Boom et al., 2017). Furthermore, the introduction of large language models (LLMs) like GPT has been pivotal in facilitating conversational recommendation scenarios. These models empower AI to generate recommendations based on natural dialogue, supporting nuanced, context-aware interactions with users that transcend mere data manipulation (Wang et al., 2022; Al-Hasan et al., 2024).

AI-driven recommendation systems also stand out due to their capability to conduct sentiment analysis, especially with regards to lyrics and user reviews. This innovative approach imbues the recommendation process with an emotional dimension often overlooked by traditional methods (Schedl et al., 2018). By integrating such sentiment analysis, AI systems can create a personalized user experience that not only aligns with surface-level preferences but also resonates with users' current emotional states or deeper interests. The emotional context of music, for example, can greatly influence user engagement, and by efficiently analyzing user sentiments, AI can better cater to individual listener moods, enhancing overall satisfaction (Li, 2024).

As AI technology continues to evolve, particularly in the realm of deep learning and natural language processing, its applications within the music recommendation sector are projected to become increasingly sophisticated. Upcoming advancements in AI will likely offer innovative approaches for comprehensively understanding and anticipating user needs, particularly in dynamic and individualized contexts (Sharma & Sharma, 2023). The proliferation of AI techniques promises to further bridge the gap between user expectations and actual recommendations, thereby creating seamless experiences that foster deeper connections between users and content.

The synthesis of advanced machine learning methodologies and natural language processing is reshaping the landscape of recommendation systems. These AI-based techniques not only refine the accuracy of suggestions but also enhance the emotional and contextual relevance of recommendations. As researchers and practitioners continue to unravel the complexities of user preferences and explore the emotional dimensions of interactions, the potential for AI in personalized recommendations remains a fertile area for academic inquiry and practical application (Schedl et al., 2018; Yang et al., 2022).

## CHAPTER TWO

### Designing Lyric-Based Music Suggestions

With a foundational understanding of recommendation system techniques established, this chapter turns to the design and development of a personalized music recommendation model centered on lyrical content. The goal of this phase is to explore how computational tools can be implemented to analyze the language of song lyrics and identify patterns that align with individual user preferences. By focusing on the textual dimension of music, this approach seeks to uncover thematic consistencies in a user's musical taste that may not be captured by traditional metadata or behavioral analysis. This chapter outlines the research questions and hypotheses that guide the model's design, details the data sources and preprocessing methods used, and describes the implementation of the TF-IDF algorithm as a mechanism for generating personalized suggestions. It also introduces the role of ChatGPT as a tool for evaluating the system's performance, setting the stage for a deeper analysis of the model's strengths and limitations.

#### *Research Question & Hypothesis*

The study is guided by two primary research questions. First, it seeks to determine how effective the TF-IDF (Term Frequency–Inverse Document Frequency) algorithm is at recommending music based solely on the analysis of lyrical content. Rather than relying on user behavior or genre-based classifications, this question addresses whether analyzing the thematic and linguistic patterns in song lyrics can meaningfully align with a

listener's personal music preferences. Second, the study investigates how ChatGPT can be utilized to evaluate and validate the quality of these recommendations. This includes exploring whether ChatGPT can effectively assess the thematic consistency, emotional relevance, and contextual accuracy of the recommended songs based on the user's explicit or implicit tastes.

The central hypothesis of the study is that the TF-IDF algorithm can successfully generate music recommendations when applied to lyrical content. The aim is for these recommendations to resonate with users once the algorithm has identified and emphasized key thematic words present in their preferred songs. It is expected that this algorithmic approach will provide a strong baseline for lyric-driven personalization. Additionally, it is hypothesized that ChatGPT can function as a reliable and insightful evaluative tool. Through its capacity to interpret and summarize patterns in lyrical data and recommendation feedback, ChatGPT is anticipated to offer meaningful judgments about the coherence and relevance of the recommended tracks. Beyond validation, its feedback is expected to reveal the strengths and limitations of the TF-IDF method.

### *Data Collection*

At the outset of our research, we identified the *Million Song Dataset* as a foundational resource for our project, which offers a robust and comprehensive collection of data on one million tracks (Bertin-Mahieux 2011). This dataset includes a wide range of relevant attributes, such as track ID, artist location, tempo, key, and even loudness, which are elements crucial for in-depth music analysis. Recognizing its potential, we chose to use this dataset not only to initiate our exploration but also to lay the groundwork for future phases of our research, where we plan to integrate these musical

features alongside lyrical content into our recommendation system. Given the extensive nature of the dataset, we invested a significant amount of time in understanding the structure and format of the h5 files and how to extract its data using the *h5py* Python library. This process involved unpacking the hierarchical data format, extracting relevant information, and organizing it into a more accessible format using Python's *pandas* library. By converting the data into a pandas DataFrame, we ensured that it was ready for efficient manipulation and analysis, which would contribute to a solid foundation for the subsequent stages of our research.

In addition to the *Million Song Dataset*, we also explored the *musiXmatch Dataset*, a large collection of song lyrics presented in a bag-of-words format that is part of the *Million Song Dataset* project. Initially, this dataset appeared to be promising due to its extensive collection of lyrical data that could be directly matched with the tracks. However, after closer examination, we decided against using the *musiXmatch Dataset*. The primary reason was that the bag-of-words format, although it is useful for certain types of analysis, did not align with our research objectives. This format represents songs as unordered collections of words, thus stripping away the contextual relationships between words and sentences. At the beginning of our research, we were uncertain about which specific methods we would ultimately employ. Ultimately, we decided we wanted to preserve the flexibility to analyze full lyrics, including sentence structures and thematic flows, which a bag-of-words approach would not support.

To address the limitations we encountered with the *musiXmatch Dataset*, we sought an alternative and discovered the *Spotify Million Song Dataset*. This dataset included the full lyrics for a million songs, which provided the depth and completeness of



data we needed for our analysis (Mahajan). Although there was not a perfect overlap between the songs in the *Million Song Dataset* and the *Spotify Million Song Dataset*, we were able to create a subset of data that successfully incorporated information from both sources. This hybrid dataset enabled us to retain all relevant musical features while also allowing for a comprehensive lyrical analysis, which includes the examination of sentence structures, thematic elements, and word usage patterns.

Once we had compiled our dataset, the next step was to refine it to ensure the quality and relevance of the data. We began by removing stopwords, which are common words that typically do not contribute meaningful information to text analysis, using the *Natural Language Toolkit (nltk)* library. This step was crucial to ensuring that our analysis focused on the most significant and impactful words in the lyrics. Additionally, we meticulously cleaned the dataset by eliminating entries with missing or incomplete information in order to further enhance our data's integrity. These preparatory steps were essential in laying the groundwork for the development of our music recommendation system and ensured that the data we used was both comprehensive and well-structured for the tasks ahead.

### *Implementation of TF-IDF*

As explained in the Background section, we employed the TF-IDF algorithm to vectorize the song lyrics by attaching weights to words and used cosine similarity to identify the most similar vectors to our current vector, which evolves based on user feedback. This iterative process allows the recommendation system to refine its suggestions dynamically. Additionally, we generated a list of “relevant words” along with their TF-IDF weights. A high TF-IDF weight for a word indicates that it is

particularly significant within the specific song's lyrics and relatively rare across the entire corpus. This distinction highlights words that are central to the song's content and differentiates it from other texts. By focusing on these high-weight terms, we enhance the system's ability to identify and recommend music that aligns closely with the user's unique preferences.

In our approach to analyzing the data, we employed an iterative methodology designed to continuously refine the music recommendation process based on user feedback. This began by presenting users with a song from our dataset. After listening, users indicated whether they liked or disliked the song. If the user expressed a preference for the song, our system used this feedback to recommend additional tracks with similar lyrical characteristics. Specifically, the system identified and suggested songs whose lyrics shared common features with the initially liked track, aiming to enhance the relevance of subsequent recommendations. Conversely, if the user did not like the presented song, the recommendation strategy would adjust accordingly. The system either randomly selected a new song from the dataset or chose a track with lyrics that exhibited distinctly different attributes from the previously shown song. This approach aimed to explore a broader range of lyrical content to better align with the user's preferences, subsequently improving the chances of finding songs that match their tastes. The core idea behind this process, in which users continuously evaluate and provide feedback on song recommendations, is that repeated interactions with the user will progressively enhance the accuracy of the recommendations. The more feedback the system receives, the better it can understand the user's preferences and refine its suggestions. The code is found below.

```

while True:
    if len(seen_indices) == len(lyrics):
        # Very unlikely to happen!
        print("No more lyrics to recommend.")
        break

    if not selected_indices:
        # If no lyrics have been liked yet, start with a random
        unseen_lyric
        unseen_indices = list(set(range(len(lyrics))) -
        seen_indices)
        current_index = np.random.choice(unseen_indices)
    else:
        # Otherwise, find the next closest lyric
        current_index, _ = find_next_closest(
            reference_vector, tfidf_matrix, selected_indices,
            seen_indices
        )

    seen_indices.add(current_index)

    print(titles[current_index], " by ", artists[current_index])
    print(f"Lyric: {lyrics[current_index]}")

    user_input = input("Do you like this lyric? (Y/N/Q):
    ").strip().upper()

    if user_input == "Q":
        break
    elif user_input == "Y":
        selected_indices.append(current_index)
        if reference_vector is None:
            reference_vector = tfidf_matrix[current_index]

```

Figure 2. Code fragment for music suggestions.

### *Performance Analysis through ChatGPT*

We then provide the output generated by the model ChatGPT 3.5, which includes a comprehensive list of all the music suggestions alongside the user's responses and indicates whether they liked each song or not. This output also captures the progression of the music recommendation system, illustrating how it refines its recommendations through iterative feedback and gradually aligning with the user's preferences.

Additionally, we include the top 10 most relevant words that the TF-IDF algorithm assigns the highest weights to, which offers deeper insight into the lyrical features that are influencing the recommendations. This process is repeated multiple times, with each iteration's output presented to ChatGPT for evaluation.

During this evaluation, we ask ChatGPT three specific questions: First, "Are these recommendations valid?" We expect a binary response to this. Second, "How confident are you in your answer?" And finally, "Please justify your reasoning." Alongside these questions, we feed the model a substantial amount of input and output data from our model to construct a rich dataset for analysis. We meticulously record ChatGPT's answers to these three key questions and store them in a document.

To further refine our understanding, we then provide this document back to ChatGPT and ask it to identify common trends and patterns in its previous responses. By doing so, we can extract the most relevant and consistent insights from ChatGPT's analysis. This allows us to not only validate the effectiveness of our music recommendation system but also to uncover strengths and weaknesses in our approach. This feedback is invaluable as it guides us in refining our model and improving the overall research.

### *Results*

For the first question, ChatGPT found the recommendations generated by the TF-IDF model to be overall satisfactory. The model was effective in identifying and recommending songs with similar lyrical content based on the keywords that were most relevant to the user's individual preferences. This indicates that the model performed well in aligning the recommendations with the user's indicated likes and dislikes. For the

second question, ChatGPT was moderately certain about the satisfactory nature of the recommendations while acknowledging that the model has some weaknesses. While the TF-IDF model succeeded in capturing important thematic words and making relevant recommendations, it was limited in its understanding of the context of these words and full alignment with the user's emotional and thematic preferences.

ChatGPT found the strengths to be that TF-IDF excels in identifying key thematic words within song lyrics and emphasizing those most relevant to user preferences. This allows TF-IDF to recommend songs with similar lyrical content and themes, contributing to a tailored listening experience. By focusing on the frequency and importance of words, TF-IDF can draw meaningful connections between songs that share common themes, making it a valuable tool for lyric-based recommendations.

ChatGPT found the weaknesses to be that TF-IDF sometimes struggles with the context in which words are used. This limitation leads to recommendations that may not fully align with the user's preferences. Although the model successfully identified keywords, it occasionally recommended songs where the word appeared but did not fit the emotional context the user preferred. This indicates that TF-IDF may overemphasize word frequency without understanding nuanced meanings within lyrics. Additionally, TF-IDF does not analyze sentiment, tone, or genre, which are important for more personalized recommendations. Thus, while TF-IDF is effective for finding similar lyrics, it might benefit from supplementary methods that consider lyrical context and emotional content.

Generally speaking, the TF-IDF model demonstrated a satisfactory performance in generating song recommendations based on lyrical content. The model effectively

identified and recommended songs with relevant themes. This suggests that TF-IDF is adept at capturing important keywords and making recommendations that reflect the user's indicated likes and dislikes. However, while the model succeeded in highlighting significant terms, it showed some blind spots in capturing the nuanced context of these words. The recommendations occasionally did not align perfectly with the emotional or thematic preferences of the user, indicating that TF-IDF might overemphasize word frequency at the expense of contextual significance. Furthermore, the lack of sentiment, tone, or genre analysis points to areas in which the model could be improved. Overall, TF-IDF is a valuable tool for lyric-based recommendations, but integrating additional methods to account for emotional context and lyrical content could enhance its effectiveness.

### *Discussion*

Our findings indicate that TF-IDF successfully identifies key words and makes relevant recommendations, improving the personalization of music discovery. This is particularly valuable in today's digital music landscape, where users face challenges in finding songs that match their specific preferences. However, our research also highlights limitations in the TF-IDF model. While it excels at pinpointing relevant terms, it tends to overlook the emotional and contextual nuances of lyrics, which can impact useful recommendations. Feedback from ChatGPT confirmed TF-IDF's effectiveness in identifying themes but also pointed out areas for improvement, such as the previously mentioned incorporation of contextual and emotional analysis.

Using ChatGPT to evaluate traditional music recommendation methods is pivotal in identifying both strengths and weaknesses in these systems. In a world where

recommendation systems are becoming increasingly integral to personalized digital experiences, the ability to critically assess and enhance these systems is vital. ChatGPT provides a sophisticated means of analysis by offering insights into the effectiveness of recommendation algorithms, such as TF-IDF, in capturing nuanced user preferences. Its feedback helps pinpoint where traditional methods excel, such as in identifying key thematic elements in lyrics, and where they fall short, such as in understanding contextual and emotional subtleties. This approach not only validates the performance of existing methods but also highlights areas for improvement, ensuring that recommendation systems can evolve to meet the growing demands for more personalized and engaging user experiences in an expanding digital landscape.

While lyrics are a significant factor in shaping our musical preferences, they are not the only elements that influence why we like certain songs. We plan to extend this research by incorporating additional attributes from the dataset, such as tempo, artist location, key, and release year. By integrating these factors, we aim to develop a more comprehensive understanding of the various influences that contribute to our music preferences. The TF-IDF algorithm was a natural choice for this research. It not only allows us to vectorize and weight lyrical content but also enables us to apply similar techniques to both quantitative attributes like tempo and release year, and qualitative ones like artist location. This approach will help us create a more nuanced recommendation system capable of explaining why certain songs resonate with listeners on multiple levels.

A promising avenue for future research involves utilizing the Recommenders library, a resource we discovered later in our study after committing to the TF-IDF algorithm. This library allows us to apply and compare multiple machine learning

algorithms within the same dataset in order to facilitate a comprehensive evaluation of their effectiveness (Recommenders 2024). Notably, the library includes advanced algorithms such as Neural Recommendation with Multi-Head Self-Attention (NRMS) introduced by Wu, Chuhan, et al. (2019), and Convolutional Sequence Embedding Recommendation (CASER) introduced by Tang and Ke (2018). NRMS leverages self-attention mechanisms to capture intricate dependencies in lyrical data that can potentially offer superior performance in understanding complex textual patterns. CASER uses convolutional neural networks to model sequential data, which could improve the system's ability to recognize and recommend based on the structure and context of song lyrics.

Additionally, the Recommenders library provides performance metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared ( $R^2$ ), which are essential for assessing the accuracy and quality of each algorithm (Recommenders, 2024). Since the library already includes TF-IDF, we can instantly compare its performance with other algorithms and obtain metrics for each run without the need for manual calculations. This feature streamlines our analysis and enhances the efficiency of our research.

While we currently use ChatGPT as our primary evaluation tool to assess the validity and strengths of our recommendation system, future studies could explore alternative AI-based evaluation methods. For instance, employing sentiment analysis models could provide insights into the emotional resonance of recommended songs, while using language models like BERT or GPT-4 for more nuanced textual analysis might offer different perspectives on lyrical content. Moreover, AI-driven systems like



reinforcement learning could be adapted to simulate user interactions and evaluate recommendation quality over time. These alternatives could complement or even enhance the insights gained from ChatGPT in order to offer a broader range of evaluative perspectives to refine and improve our recommendation system.

## CHAPTER THREE

### Evaluating AI's Capabilities as a Music Recommender

In Chapter Two, we explored the effectiveness of traditional content-based filtering algorithms, specifically the TF-IDF model, and utilized ChatGPT as an evaluative tool to analyze the model's ability to recommend music based solely on lyrical content. While this computational approach offered valuable insights into thematic similarities between songs, the analysis highlighted significant limitations regarding the model's inability to capture nuanced emotional contexts and deeper user preferences fully. Acknowledging these limitations, Chapters Three and Four of our research seeks to address these gaps through direct user interaction. By employing an interactive survey approach, we strive to validate and enhance our computational findings empirically. The purpose of this survey is to leverage real-world user feedback to evaluate ChatGPT's effectiveness in iterative music recommendation scenarios and refine our understanding of personalized AI-driven recommendations.

#### *Research Question & Hypothesis*

The second phase of this study shifts the focus toward user interaction and the dynamic refinement of AI-generated music recommendations. It is guided by two key research questions. First: how effectively can ChatGPT iteratively refine its music recommendations based on direct binary user feedback (i.e., like or dislike)? This question explores the adaptive capacity of a conversational AI model when engaged in repeated recommendation rounds. We seek to understand whether such a system can

progressively learn from minimal input and tailor its suggestions to better match user preferences over time. Second: what is the extent to which users perceive ChatGPT's final analysis of their musical preferences as both accurate and insightful? This involves evaluating not just the song suggestions themselves, but also the AI's ability to synthesize patterns and present meaningful summaries of individual taste profiles.

The hypothesis underpinning this phase is: when guided by iterative binary feedback, ChatGPT will progressively improve the accuracy and personalization of its music recommendations. It is expected that users will respond positively to this refinement process, as the system becomes increasingly attuned to their preferences with each round of interaction. Furthermore, we hypothesize that users will generally find ChatGPT's final analyses to be accurate reflections of their musical tastes. These insights may include genre preferences, lyrical themes, tempo tendencies, and emotional tone, which would demonstrate ChatGPT's potential not only as a recommendation tool but also as a perceptive and explanatory AI capable of capturing the nuances of individual music identities.

### *Purpose of the Survey*

The interactive survey approach is essential to supplement and validate computational evaluations from Part 1 with empirical evidence gathered directly from human interactions. While computational algorithms like TF-IDF effectively identify thematic similarities based on textual content, human preferences in music involve complex emotional, contextual, and experiential dimensions that algorithms alone tend to overlook. Thus, incorporating user feedback provides crucial insights into the practical performance of recommendation systems, which will bridge the gap between theoretical

assessments and real-world applicability. The feedback collected from participants will allow us to critically examine whether iterative interactions between humans and AI enhance the relevance and personalization of music recommendations.

### *Survey Design*

It starts with a specific prompt that we feed to the model (see Figure 3). Then, the survey is structured as a 10-round iterative interaction between users and ChatGPT. This approach allows ChatGPT to recommend a sequence of songs while incrementally refining its suggestions based on the participant's binary feedback, which simplifies the user's response to a straightforward indication of liking or disliking each song. The rationale for using binary feedback lies in its clarity and efficiency. It quickly guides the AI toward a clearer understanding of individual user preferences. At the end of the 10 rounds, ChatGPT provides a comprehensive analysis summarizing the participant's musical preferences, capturing attributes like genre, tempo, lyrical themes, key, and other relevant musical or contextual features. Participants subsequently rate the accuracy and quality of ChatGPT's analysis, providing qualitative feedback that further enriches our evaluation. The survey instrument and the research protocol were reviewed and approved by Baylor's IRB with #2270562.

Please, open ChatGPT and copy and paste this code into chatGPT (free version):

Please recommend a random song that appeared on the Top 100 charts in the United States between 1990 and 2024. I will listen to it and provide feedback by indicating whether I like it or not. Based on my feedback, recommend another song, aiming to refine your understanding of my musical preferences. After 10 iterations, generate a detailed report summarizing my music preferences and dislikes. In your analysis, consider attributes such as genre, tempo, key, lyrical themes, and any other relevant musical or contextual features. Ensure the report is as comprehensive and insightful as possible, but only report this information after the 10 iterations are done.

Figure 3.ChatGPT Prompt

### *Participant Criteria and Recruitment*

To ensure reliable and meaningful results, the survey establishes clear participant eligibility criteria. Participants must be at least 18 years old, proficient in English, have access to an online music streaming platform (such as Spotify or Apple Music), and demonstrate willingness to engage with ChatGPT through the 10-round iterative recommendation session. These criteria ensure participants can understand the recommendations, interact effectively with the AI system, and provide clear, relevant feedback. The recruitment process aims to capture diverse musical tastes and demographic backgrounds, thereby enhancing the generalizability and applicability of the findings.

### *Procedure of the Interactive Session*

Participants are instructed to initiate their interaction with ChatGPT by entering a specifically formulated prompt, asking ChatGPT to recommend a random song from the U.S. Top 100 charts between 1990 and 2024. After listening to each recommendation, participants indicate whether they liked or disliked the song. This binary response informs ChatGPT's subsequent recommendation, which enables it to iteratively adapt and

refine its understanding of the user's musical preferences across the 10 iterations. Upon completion of this iterative recommendation process, ChatGPT generates a detailed analytical report summarizing the participant's overall music taste. Participants then copy this report into the survey, evaluate its accuracy, and rate their satisfaction as well as a final qualitative justification for their rating. This structured approach allows for a systematic evaluation of the AI's ability to understand and align with user preferences over time progressively.

## CHAPTER FOUR

### Results and Reflections from User Interaction

Following the design and execution of an interactive survey using ChatGPT as a music recommender, this chapter presents the results of the user study and reflects on its real-world implications. The aim of this phase was not only to assess the AI's ability to refine recommendations over time but also to understand how users perceive and respond to personalized suggestions generated through conversational interaction. This chapter analyzes user satisfaction, perceived accuracy, and the emotional and contextual relevance of the recommendations by drawing from respondents' quantitative responses and qualitative feedback. It also considers recurring challenges and user-reported limitations that allow us to understand the current capabilities of AI-driven personalization. These insights provide a valuable bridge between computational design and real-world application, which will help us determine any necessary future improvements in recommendation systems.

#### *Analysis and Results*

Participants exhibited relatively high initial satisfaction, with approximately 72.2% positively rating the first music recommendation provided by ChatGPT. Throughout the iterative recommendation rounds, satisfaction remained generally high, reaching its peak in the final, tenth iteration, where 88.9% of participants expressed liking the recommended song. The lowest satisfaction was observed during the fourth recommendation round, with only 58.3% of participants responding positively.

Participants evaluated the accuracy of ChatGPT's analysis of their music preferences with an average rating of 7.43 out of 10, indicating a fairly strong perception of accuracy (refer to Figure 4). Furthermore, when explicitly asked about the accuracy of ChatGPT's final summary, a substantial majority of participants (30 out of 36) affirmed that the analysis was indeed accurate, while the remaining participants indicated dissatisfaction or perceived inaccuracies.

Qualitative feedback from participants offered deeper insights into their quantitative ratings. Common positive feedback highlighted ChatGPT's effectiveness in identifying significant musical elements such as instrumentation, tempo, and genre. Participants often noted that ChatGPT accurately captured their general musical preferences and accurately distinguished between liked and disliked musical attributes, which demonstrates effectiveness in its iterative refinement.

However, participants also mentioned certain limitations. Many perceived ChatGPT's recommendations as overly narrow or excessively focused on mainstream or popular music. This led to repetitive suggestions within a limited set of artists or genres once the participant indicated their initial preferences. Furthermore, ChatGPT seemed to struggle when addressing niche or less conventional music preferences, such as punk, goth, darkwave, or country genres. Several participants felt the system became biased by early positive feedback, which hindered further exploration. Additionally, participants indicated that ChatGPT often overlooked contextual and emotional dimensions of musical experiences, such as listener mood and familiarity with certain songs. The binary feedback mechanism was also consistently identified as a limitation. Participants



expressed that this format limited their ability to convey nuanced preferences, which affected the depth and accuracy of subsequent recommendations.

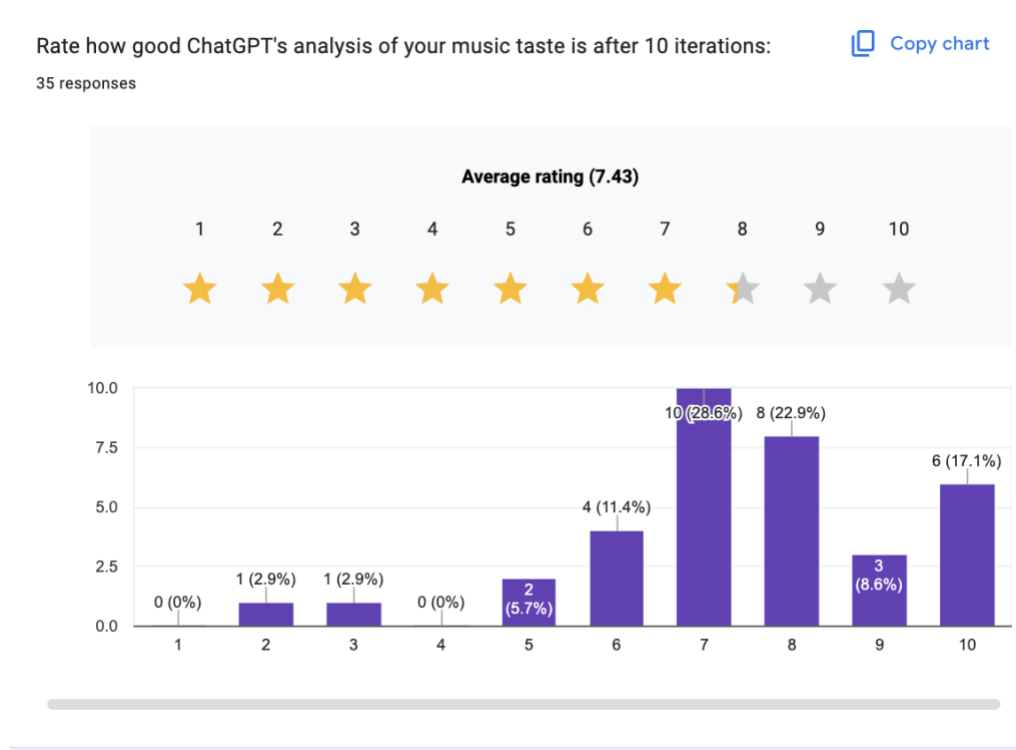


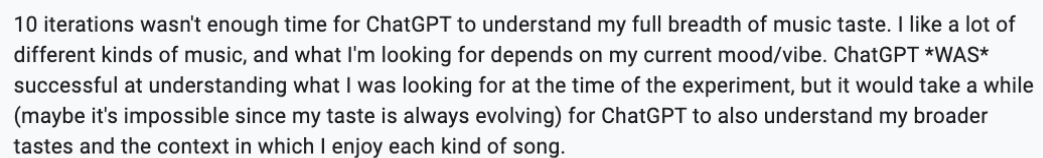
Figure 4. "Rate how good ChatGPT's analysis of your music taste is after 10 iterations"

### Discussion

The findings obtained from this user-interactive survey provide critical empirical validation regarding ChatGPT's capability to refine music recommendations based on direct user feedback. High overall satisfaction rates, especially by the tenth iteration, indicate that ChatGPT is capable of progressively aligning its recommendations with individual user preferences. Moreover, the positive participant evaluations confirm that ChatGPT can accurately identify key musical attributes, validating and complementing computational insights from Part 1.

Despite the overall positive outcomes, significant limitations emerged from user feedback. Participants frequently commented on the limited musical diversity offered by ChatGPT, which appeared constrained by its reliance on more mainstream tracks (refer to Figure 5). This limitation was particularly noticeable among participants with more niche music tastes, who found ChatGPT’s iterative process overly narrow or repetitive. These concerns echo earlier computational findings that indicated difficulties in fully capturing nuanced emotional and thematic contexts. Participants’ consistent highlighting of the binary feedback limitation further reinforced this challenge and emphasized the necessity for more sophisticated mechanisms that can accommodate nuanced preferences.

The integration of human interaction thus proved invaluable in revealing critical aspects of music preference, such as emotional context and broader genre preferences, which computational analyses alone may fail to capture adequately. The empirical results underscore the importance of user engagement and the potential for interactive feedback to enhance AI-driven recommendation systems.

A screenshot of a user's response to the question "Explain your rating". The text is displayed in a light gray box with a thin border. The user's response is: "10 iterations wasn't enough time for ChatGPT to understand my full breadth of music taste. I like a lot of different kinds of music, and what I'm looking for depends on my current mood/vibe. ChatGPT \*WAS\* successful at understanding what I was looking for at the time of the experiment, but it would take a while (maybe it's impossible since my taste is always evolving) for ChatGPT to also understand my broader tastes and the context in which I enjoy each kind of song."

10 iterations wasn't enough time for ChatGPT to understand my full breadth of music taste. I like a lot of different kinds of music, and what I'm looking for depends on my current mood/vibe. ChatGPT \*WAS\* successful at understanding what I was looking for at the time of the experiment, but it would take a while (maybe it's impossible since my taste is always evolving) for ChatGPT to also understand my broader tastes and the context in which I enjoy each kind of song.

Figure 5. Response by one user to the question of “Explain your rating”

### *Future Work and Recommendations*

Based on these insights, there are several recommendations for future research and development to address current limitations and strengthen music recommendation

systems. Expanding the scope and diversity of music datasets beyond mainstream charts will enable ChatGPT to better cater to varied and niche user preferences. Also, adopting more sophisticated feedback mechanisms (e.g. Likert scales, qualitative textual feedback, or multi-dimensional rating systems) would allow users to convey more detailed preferences in order to facilitate deeper personalization.

Conducting longitudinal studies could also help determine whether prolonged interactions significantly improve recommendation accuracy, as well as reveal long-term trends and adaptations in user preferences. Real-time sentiment or emotional analysis during listening sessions may further enhance recommendation context, which would provide richer emotional and experiential data to complement traditional analytical dimensions.

Moreover, exploring advanced AI methodologies like reinforcement learning could dynamically adjust recommendations beyond the constraints of simple binary feedback. Comparative studies involving advanced language models such as GPT-4 or BERT would also help benchmark system performance and potentially improve future iterations. Expanding participant diversity and increasing sample sizes are further recommended to enhance the generalizability and robustness of future findings across various user populations.

## CHAPTER FIVE

### Conclusion

As this study concludes, it is important to revisit the two interconnected phases that shaped the research. Part 1 focused on the development of a lyric-based recommendation system using the TF-IDF algorithm. We evaluated its ability to identify thematic patterns and generate music suggestions aligned with user preferences. This computational phase provided a structured foundation for understanding how content-based filtering can be applied to song lyrics. Part 2 expanded on this by introducing a user-interaction component, where participants engaged in iterative recommendation sessions with ChatGPT to assess the relevance of AI-generated suggestions. Together, these two parts illustrate the value of combining algorithmic modeling with empirical user feedback. The subsequent sections offered reflections on each phase by highlighting key findings, limitations, and opportunities for future development in AI-powered music recommendation systems.

The first phase of this research aimed to evaluate the effectiveness of content-based recommendation techniques by leveraging the TF-IDF algorithm to analyze lyrical content. This approach offered a focused, interpretable framework for identifying thematic consistencies in users' musical preferences. By converting song lyrics into numerical vectors and applying cosine similarity, the system successfully generated recommendations aligned with users' explicit or implicit tastes. The integration of ChatGPT as an evaluative tool further enriched the study and allowed for detailed

feedback on the thematic relevance and coherence of the generated recommendations. Overall, the results affirmed that, when structured through TF-IDF, lyrical analysis can produce personalized and logically grounded music suggestions.

Despite these strengths, the model displayed several limitations that revealed the need for further development. While it performed well in identifying frequently used thematic words, it often struggled to account for emotional nuance, lyrical context, and deeper semantic meanings. Songs with similar keyword profiles sometimes differed dramatically in tone or message, which contributed to occasional mismatches in recommendations. Moreover, the algorithm's sole reliance on word frequency limited its ability to consider non-lyrical aspects of music, such as genre, instrumentation, or cultural context. These limitations pointed to the importance of integrating broader forms of analysis and feedback into the system.

These findings underscore the value of combining algorithmic precision with evaluative intelligence. While TF-IDF provides a solid baseline for recommendation based on lyrical similarity, it falls short in fully capturing the subjective and emotional dimensions of musical experience. This realization laid the groundwork for the second phase of the study, where direct user interaction was introduced to address the limitations detailed in Part 1. By complementing computational modeling with human input, this research seeks to provide a roadmap for a more comprehensive and responsive music recommendation system. Specifically, one that goes beyond surface-level analysis to engage with the full complexity of individual taste.

The second phase of the study offered critical insights into the role of user interaction in refining AI-generated music recommendations. Through ten rounds of

feedback, participants engaged in a dynamic process where their input directly influenced the AI's understanding of their musical tastes. The findings showed that iterative feedback led to progressively more accurate and satisfying recommendations, thereby validating the computational results from Part 1. Users generally affirmed the accuracy of ChatGPT's final summaries of their preferences, underscoring the model's potential to synthesize meaningful patterns from minimal input.

However, the results also brought important challenges to light. Participants consistently pointed out the narrow range of genres offered, the repetitive nature of suggestions, and the system's limited ability to understand or reflect emotional and situational factors in their music choices. These shortcomings were especially apparent among users with unconventional preferences. The binary feedback mechanism, while simple and user-friendly, was also seen as too restrictive to convey the complexity of an individual's musical taste. These findings suggest that even advanced AI models benefit significantly from richer, more diverse forms of user input.

The integration of computational modeling with empirical user evaluation proved to be a powerful combination. It allowed for a more holistic understanding of how personalization in music recommendation can be both designed and experienced. Future research should explore more flexible feedback systems (e.g. scaled ratings, emotion tagging, or natural language input) and try to diversify the training datasets to better accommodate underrepresented genres and listener profiles. Incorporating additional AI techniques like sentiment analysis, reinforcement learning, or multimodal recommendation models could also strengthen the system's ability to respond to emotional and contextual cues.

Ultimately, this study demonstrates that while algorithms like TF-IDF offer a strong starting point for personalized music recommendations, they reach their full potential when combined with human insight and interaction. AI systems are most effective not when they replace user preferences, but when they learn from them. By incorporating the complexity of human experience with algorithmic efficiency, we can move closer to building recommendation systems that are not only accurate, but also emotionally intelligent, inclusive, and genuinely meaningful to the people they serve.

## BIBLIOGRAPHY

- Al-Hasan, T., Sayed, A., Bensaali, F., Himeur, Y., Varlamis, I., & Dimitrakopoulos, G. (2024). *From traditional recommender systems to gpt-based chatbots: a survey of recent developments and future directions*. Big Data and Cognitive Computing, 8(4), 36. <https://doi.org/10.3390/bdcc8040036>
- Bertin-Mahieux, Thierry, et al. "The Million Song Dataset." Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.
- Boom, C., Agrawal, R., Hansen, S., Kumar, E., Yon, R., Chen, C., ... & Dhoedt, B. (2017). *Large-scale user modeling with recurrent neural networks for music discovery on multiple time scales*. Multimedia Tools and Applications, 77(12), 15385-15407. <https://doi.org/10.1007/s11042-017-5121-z>
- Goodfellow, Ian, et al. *Deep Learning*. MITP, 2018.
- Li, J. (2024). *Research on music neighborhood-based recommendation algorithms*. Applied and Computational Engineering, 111(1), 124-130. <https://doi.org/10.54254/2755-2721/111/2024ch0098>
- Lokesh, Ashwini, "A Comparative Study of Recommendation Systems" (2019). Masters Theses & Specialist Projects. Paper 3166. <https://digitalcommons.wku.edu/theses/3166>
- Mahajan, Shrirang. "Spotify Million Song Dataset." Kaggle, 21 Nov. 2022, [www.kaggle.com/datasets/notshrirang/spotify-million-song-dataset](https://www.kaggle.com/datasets/notshrirang/spotify-million-song-dataset).
- Rajaraman, Anand, et al. *Mining of Massive Datasets*. Cambridge University Press, 2014.
- Recommenders Team. *Recommenders*. GitHub, 2024, <https://github.com/recommenders-team/recommenders>.
- Schedl, M. (2019). *Deep learning in music recommendation systems*. Frontiers in Applied Mathematics and Statistics, 5. <https://doi.org/10.3389/fams.2019.00044>
- Schedl, M., Zamani, H., Chen, C., Deldjoo, Y., & Elahi, M. (2018). *Current challenges and visions in music recommender systems research*. International Journal of Multimedia Information Retrieval, 7(2), 95-116. <https://doi.org/10.1007/s13735-018-0154-2>



- Sharma, A. and Sharma, R. (2023). *The role of generative pre-trained transformers (gpts) in revolutionising digital marketing: a conceptual model*. JCMS, 8(1), 80. <https://doi.org/10.69554/tlvq2275>
- Tang, Jiayi, and Ke Wang. "Personalized Top-N Sequential Recommendation via Convolutional Sequence Embedding." Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, 2018.
- Wang, X., Zhou, K., Wen, J., & Zhao, W. (2022). *Towards unified conversational recommender systems via knowledge-enhanced prompt learning*. <https://doi.org/10.48550/arxiv.2206.09363>
- Wu, Chuhan, et al. "Neural News Recommendation with Multi-Head Self-Attention." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.
- Yang, X., Chen, A., PourNejatian, N., Shin, H., Smith, K., Parisien, C., ... & Wu, Y. (2022). *A large language model for electronic health records*. NPJ Digital Medicine, 5(1). <https://doi.org/10.1038/s41746-022-00742-2>