Addressing Dataset Scarcity in Music Emotion Recognition with LLMs

Janita Aamir and Patrick J. Donnelly[0000−0003−1033−5931]

Oregon State University, Corvallis OR 97330, USA
{aamirj,donnellp}@oregonstate.edu

Abstract. Music Emotion Recognition is an important task in the field of music information retrieval, with applications in music generation, song recommendation, and playlist generation. However, given copyright concerns and the expense of human subject studies, researchers lack large datasets annotated with emotions. This dataset scarcity significantly limits the ability of researchers to train machine learning models that recognize emotion directly from audio. To address this need, we introduce a novel approach that leverages language models to generate emotion annotations. Specifically, we task ChatGPT to provide (1) numeric estimations (2) set of emotion words, and (3) long-form descriptions that characterize the emotive qualities of a piece of music. We consider 22,968 songs across five public datasets to facilitate a comparison of our ChatGPT synthetic annotations against human and previous algorithmic annotations. Although indirect, these annotations have the potential to provide insight into the emotional content of music, opening new possibilities for research and applications in the area of music emotion recognition.

Keywords: Music Emotion Recognition · Language Models · ChatGPT · Dataset · Annotations · Circumplex Model · Valence · Arousal

1 Introduction

Music has the power to elicit strong emotional reactions in listeners. The ability to predict these emotions directly from music itself is a task of great interest to researchers in the field of music information retrieval. Music Emotion Recognition (MER) is the process of estimating the emotional content of a piece of music using computational methods, such as machine learning. If researchers could automatically predict common emotional reactions to music directly from the analysis of acoustic signals alone, this would open up new possibilities for music recommendation systems, mood-based playlist generation, and music therapy.

Prior to the rise of deep learning, researchers primarily relied on signal processing to extract features for predicting emotional content directly from audio signals. These include high-level features, such as tempo, pitch, and timbre, as well as low-level features, such as harmony, melody, and rhythm [23]. However, despite advances in machine learning techniques, researchers report limits in their ability to predict emotional content based on audio descriptors alone [22].

The single greatest challenge faced by researchers in MER is the scarcity of large, diverse, and reliable datasets. The task of annotating music with emotion labels is resource-intensive as it requires expensive and time-consuming human subject studies. Such studies take place in a controlled laboratory setting or online in a crowd-sourced task. Despite attempts to train annotators, inter-rater agreement on perceived emotions is often quite low [27,32]. Furthermore, copyright concerns limit the ability of researchers to release large public datasets. Together these factors hinder the creation of datasets annotated with emotions.

Nevertheless, researchers have released a few annotated datasets for MER. However, these datasets are small, homogeneous, or copyrighted (see Section 3). Other researchers have explored folksonomy representations, such as using a proxy of LastFM[1] tags [18], as a means to estimate emotional responses to music from crowd-sourced text labels. Other approaches attempted to estimate emotion based on lyrics [5], social media discourse [2], and historical documents [16].

This paucity of large datasets labeled with emotion reactions to music hinders the ability of researchers to develop robust MER systems that generalize across different genres, languages, and cultures. State-of-the-art deep learning approaches require datasets that are orders of magnitude larger than any existing datasets. Because it is infeasible to curate human annotations on the scale required, researchers could potentially investigate creative computational approaches. The recent availability of large language models (LLMs), such as ChatGPT[2], has revolutionized many tasks. These models have been trained on a wide array of textual information including song titles, lyrics, music descriptions, tags, and reviews, and hold the potential to estimate the emotional content of songs.

We propose a new approach to creating datasets for the field of MER using text-based prompts to harvest valence and arousal values from a LLM. By leveraging existing LLMs, and bypassing the expense of manual annotation, there is potential to create large datasets necessary to train robust, accurate MER models. Although human annotations will always be preferred, synthetic estimations are able to produce large-scale datasets at a substantially lower cost.

Since the primary focus of this study is to explore the validity of our approach, we focus on songs that have previously existing human or algorithmic annotations. We obtain estimations on five commonly used datasets in MER: AMG1608 [6], DEAM [1], DEAP [17], Deezer [7], and PMEmo [33]. For each song within the five datasets, we use GPT-3.5-turbo to estimate annotations using three different approaches: (1) direct estimations of arousal and valence, (2) five descriptive words that represent the emotional content of the song, and (3) long-form textual descriptions that summarize the emotions conveyed. In the second approach, each word, or its synonym, is matched in a word list that contains the arousal and valence values [31]. In the third approach, the long-form text is filtered to remove stop words, and then each word from the paragraph, or its synonym, is referenced from the same list. These individual values are averaged to produce an

[1] https://www.last.fm/ (last accessed on 2026/01/20)
[2] https://chat.openai.com/ (last accessed on 2026/01/20)

estimate of arousal and valence for that song. Through these three approaches, we generated arousal and valence estimates for a total of 22,968 songs.

This novel approach of annotating music for MER tasks unleashes creative possibilities to build large-scale datasets without extensive resources. By testing this method, we aim to lay the foundation for efficient and scalable MER dataset creation. While this study focuses on validating this approach, future work will explore diverse musical genres and cultures. With the release of larger datasets, the field of MER has potential for transformative growth and new applications.

2 Background and Related Work

Researchers in MER seek to correlate characteristics of music with both perceived and induced emotions. However, to do so effectively, researchers require knowledge of which emotions are associated with which music. The data creation process either requires listeners to manually annotate emotions in music or computational processes that synthetically estimate the common emotional reactions to music.

2.1 Types of Annotations

Datasets used in music emotion recognition tasks are most often annotated at one of two temporal levels, using either static or dynamic labels. Static annotations are given at the song level in which a single label serves as a proxy for the emotions of the entire song. Although common, static annotations fail to capture emotional variation within a song. As such, they are not particularly well suited for certain genres or longer pieces of music. Dynamic annotations are marked at fixed intervals to track emotional changes over time. While the use of dynamic labels may better capture the variable nature of emotion within a piece of music, collecting dynamic labels from human subjects presents a number of additional challenges compared to the simpler task of static annotations [1].

In addition to the temporal level of annotation, datasets in MER are labeled using one of two different concepts: categorical [12] or dimensional [26] models of emotion. Categorical labels reflect a finite set of basic emotion categories, such as anger, disgust, fear, happiness, sadness, and surprise [8]. In the widely used circumplex model of dimensional encoding [25], emotions are categorized in a two-dimensional space of *arousal* (calm to exciting) and *valence* (negative to positive). The arousal dimension attempts to measure wakefulness versus sleepiness while the valence dimension attempts to measure emotions along an axis of pleasure versus displeasure. Additionally, many studies consider a discretization of the arousal and valence space, considering only the four quadrants of the space.

2.2 Limitations and Criticisms

In many domains, datasets are prepared for supervised machine learning tasks by enlisting domain experts to provide annotations. However, these individuals may differ in opinion, leading to inconsistencies in annotations [29,32]. Furthermore,

annotators in crowd-sourcing tasks experience fatigue which can negatively affect their efficiency and accuracy [34]. This noisy labeled data presents great challenges to researchers training machine learning models. The difficulty of curating high-quality annotations for MER tasks is compounded by the highly subjective nature of emotional reactions to music [19]. Emotions elicited from music often vary between individuals, given broad differences in personal and socio-cultural preferences. Although some characteristics of the music itself may correlate with perceived emotions, researchers also find differences in these perceived emotions depending on demographics or musical experience [27]. Furthermore, there are distinctions between perceived and induced emotions (see [10] for a review).

For these reasons, the MER task of annotating music with emotion has been widely criticized [11]. There exist many differences in the methodologies used to create datasets for MER. Although probabilistic approaches to map between categorical and dimensional labels have been proposed [30], researchers observed little generalizability between datasets when the method of creation differed [14].

3 Emotion Datasets

In this section we describe the five datasets we consider, review the original data collection methods, and discuss limitations and biases of each approach.

3.1 AMG1608

The AMG1608 dataset [6] consists of 1608 songs that have been annotated by 665 subjects (320 female; 345 male). The dataset provides 30-second audio clips for each song. The music is primarily contemporary Western, as categorized by the All Music Guide[3]. As part of their annotation collection process, the authors divide the 1608 songs into 134 different groups, attempting to provide a balanced mixture of songs with different emotions. Each subject is tasked to annotate a set of 13 songs in a group, which takes an average of 455 seconds to complete.

Despite the rigor of the study design, there exist some limitations in the data collection process. First, the authors only use a single genre of music, which likely introduces a cultural bias, limiting the generalization of the dataset to other genres. They also limit themselves to only 30-second audio previews of each song that may not fully represent the full range of emotions that could evolve over the course of the entire song. For their study, the authors recruited the majority of the subjects from Amazon Mechanical Turk, who were required to be residents of the United States with a high acceptance rate on that platform. This choice introduces a cultural and geographic bias, because the perception and interpretation of music may vary across different cultures and regions. Although the authors recruit 22 subjects from a campus setting, this small sample adds to potential bias because they likely share geographic and cultural characteristics.

Moreover, in a single session, subjects were tasked to annotate 13 songs within an annotation group. This number of consecutive annotations could increase the

[3] https://www.allmusic.com/ (last accessed on 2026/01/20)

risk of annotation fatigue and affect the annotation quality, especially towards the end of each session. Although the authors do not detail incentives provided to the subjects for task completion, the quality of annotations can be impacted significantly depending on the subjects' motivation and the value of the incentive.

3.2 DEAM

The DEAM dataset [1] consists of annotations for 1802 songs (58 full-length songs and 1744 45-second excerpts) collected over three years. The dataset features songs from a variety of Western popular music genres, such as rock, pop, electronic, country, and jazz. These songs were chosen because they are royalty-free, allowing the researchers to release excerpts of the songs with their dataset. However, the lack of demographic information from annotators is a significant oversight in their methodology. Demographic factors such as age, gender, country of origin, and music knowledge can significantly influence emotional responses to music. Without this, it is challenging to assess how representative the annotations are of a broader population or to understand the potential biases in the dataset.

Most of the audio in the dataset is 45-second music excerpts extracted from random moments in the song. Although this method aims for fairness, the selected excerpts might not accurately capture the emotional content of the songs because those 45 seconds may not be representative of the complete song. The authors also found that participants' preferences for certain songs influenced self-reported confidence in rating but they did not provide a clear analysis of how these factors might have influenced the actual annotations in their dataset.

3.3 DEAP

The DEAP dataset [17] is a multimodal dataset analyzing human affective states. The authors recorded the electroencephalogram and peripheral physiological signals of 32 (50% female; 50% male) participants. These participants were tasked to watch 40 music videos that were each one-minute long. These music videos were selected in a semi-automatic and manual approach in an attempt to reduce selection bias. In the semi-automatic approach, the authors made use of a list of stimuli values to retrieve songs with corresponding LastFM tags. They made sure that the songs they selected were equally spread out across the quadrants to reduce bias towards certain emotions. The authors also manually picked the songs that had the clear emotional content and were equally spread out on the arousal-valence space. The design choice of a one-minute duration for each music video was also selected by using a linear regression approach [28].

For each music video, participants provided ratings of arousal, valence, like/dislike, dominance, and familiarity. Frontal face videos were also recorded for 22 of the 32 participants. The authors highlight that during the collection process, participants were able to stop annotating, which helped reduce annotation fatigue, potentially leading to more reliable results. Overall, the authors attempted to conduct an annotation strategy that aimed to reduce bias. The inclusion of ratings such as like/dislike and familiarity could perhaps be beneficial to identify bias.

However, the study lacks a demographic breakdown of participants and more information on the participant selection process, as the authors only provided the gender and age range. Although it may be possible that demographic statistics were intentionally withheld from publication, given the small sample size.

3.4 Deezer

The Deezer mood detection dataset [7] is a large and algorithmically annotated dataset. This dataset is one of the largest available, consisting of synthetic annotations for 18,644 songs. The arousal and valence values were calculated by considering mood-related LastFM tags using the strategy proposed by Hu et al. [13]. The extracted tags are then referenced against a word list [31] to find the arousal and valence value of each tag. If several tags are present for one song, the authors take the average of the respective values. These arousal and valence values are then normalized to [0,1] to reduce the scale of the annotated values.

Although the audio and lyrics of the songs are not included in this dataset, the authors provide a mapping to the Million Song Dataset [3] so researchers can reference audio features and lyrics. The authors also do not provide any details on the average number of tags considered for each song. However, because there are only 2762 unique points across the 18,644 songs, it can be assumed that many songs share a small number of LastFM tags. Furthermore, the study did not provide details on their song selection process and the distribution of data across different genres of music. Even though human annotations are generally preferred in MER tasks, this dataset has received attention for its size and novelty.

3.5 PMEmo

The PMEmo dataset [33] contains annotations of 794 songs as well as electrodermal activity signals of the participants listening to the music. The authors recruited 457 subjects in total (236 female; 221 male), of which 366 were Chinese university students who are non-music majors, 44 were Chinese university students who are music majors, and 47 were identified as native English speakers. Each song in their dataset received a total of at least 10 emotion annotations including one by a music major and one by an English speaker. In the dataset, the authors mainly targeted trending pop songs on the Billboard Hot 100, the iTunes Top 100 Songs (USA), and the Top 40 Singles Chart (UK).

Because the majority of the annotators were Chinese students with non-music backgrounds, it is likely that there are age, academic, and cultural biases in the dataset. Although the dataset focuses on English pop songs, only 47 out of the 457 subjects (10.3%) were English speakers. This potentially introduces a significant linguistic bias, as many of the non-English speakers were unlikely to fully comprehend the nuanced emotions expressed through the English language lyrics. All these factors potentially affect the generalizability of the arousal and valence annotations collected for this dataset.

Additionally, the exclusive use of trending pop songs limits the diversity of the dataset. This focus, on English-language pop songs, might skew the balance

of arousal and valence annotations towards one of four quadrants, which is problematic as it can overlook the broader spectrum of emotions. The authors even acknowledge that a large proportion of the songs received annotations that place them in the first quadrant, the quadrant that contains high energy songs.

The paper highlights that the facilitators created a relaxing environment for the subjects before collecting annotations. During this period, subjects listen to the song "In My Life" by Kevin Kern. This requirement to listen to one particular song is problematic and introduces bias that may affect the annotations for subsequent songs. The authors also report providing a 10-second break between each annotation. Although a break is likely a good idea, the authors provided no evidence that 10 seconds was sufficient for an emotional reset for the annotators.

While annotating, subjects were required to listen to each song twice in order to rate arousal and valence separately. This design choice could lead to fatigue or decreased attentiveness, potentially affecting the reliability of annotations. When collecting annotations for their dataset, the authors only consider the chorus of the song. This introduces yet another source of bias since the annotations might not reflect the full range of emotions expressed. In fact, the chorus is often more high-energy than other parts of a song. This choice likely influenced the dataset, skewing it towards capturing more intense emotional responses.

3.6 Summary and Implications

From the analysis of each data collection study, it is clear there is a lot of variation in the annotation approaches across the datasets. This variation includes the song selection process, setting, demographics, and methods of collection. These factors affect the consistency and the reliability of the emotion labels. For example, DEAP's controlled lab-based setting results in more consistent and reliable annotations but limits participation and dataset size. In contrast, the online methods used by PMEmo and DEAM increase the dataset and participant size but tend to reduce annotation reliability. It is not feasible to create large-scale datasets through human annotations alone and, given the inconsistencies in annotation collection studies, combining smaller datasets for MER model training becomes difficult as well. This highlights the need for a standardized and scalable approach to create reliable annotations across different types of music.

4 Dataset Creation Process

Our goal is to estimate the arousal and valence values of the songs that exist in published human-annotated datasets used by the MER field: AMG1608 (n=1608 songs), DEAM (n=1802), DEAP (n=120), and PMEmo (n=794). By estimating the emotions of the songs that exist in these datasets, we are able to compare the annotations estimated through LLMs to human-annotated baselines. We also compare to synthetic annotations of the Deezer dataset (n=18,644). As one of the largest and most widely used datasets, it is an important baseline against which we compare our approach. In this study, we task GPT-3.5-turbo to obtain arousal

Table 1. Example prompt given to ChatGPT for the 5-word approach.

Example Prompt

Task:
Analyze the song `{name}` by `{artist}` to identify the emotional content conveyed in the song. You can look at lyrics, analyses, reviews, interviews, fan discussions, and more. Based on your analysis, provide five descriptive words that accurately capture the song's emotions.
 	
Output:
Your output must be strictly in readable `JSON` format without any extra text:
`{`
`"words":`
`[word1, word2, word3, word4, word5],`
`"reasons":`
`[reason for word1, reason for word2, reason for word3,`
`reason for word4, reason for word5]`
`}`
 	
Requirements:
For the 'outputs' field, enter the words that accurately describe the emotional content of the song. No words should be repeated for the song.
 	
For the 'reasons' field, provide reasoning for each word separately. The reason for each word should be in the same order as the word.
 	
If you do not know the answer to a question, do not share false information. Instead, you should enter `None` for the 'words' and 'reasons' field.

and valence values using three different methods: (1) direct, numeric estimate, (2) list of five descriptive words, and a (3) long-form textual description.

We ask ChatGPT for the emotional content of songs, following the structured approach by Zhuo et al. [35]. An example prompt for one of the methods is provided in Table 1. The prompts for the other two methods are very similar.

We structured our prompts to minimize instances of hallucination in the responses provided by ChatGPT. Hallucination occurs when the text seems sensical, but the information provided is inaccurate or invented [15]. Following the approach of Zhuo et al. [35], we found reduced instances of hallucination in the responses provided by ChatGPT. The prompt uses specific task descriptions which reduces ambiguity and prevents the model from misinterpreting the objective. A structured output format, required by the prompt, also helps constrain the model for more consistency. The requirement of a clear JSON format helps prevent the language model from generating content that is disorganized or irrelevant. In the prompt, the language model is also instructed to respond `None` if unsure about a response, which prevented the model from fabricating information. In the early stages of testing, we found that without this instruction, ChatGPT was more prone to providing responses that were inaccurate. We also instructed ChatGPT to provide reasoning with each response. This enabled the model to self-reflect on the responses and also helped us understand the decision making process of the LLM. The reasoning became especially useful when we were manually reviewing the responses. Lastly, we made sure to assign a task specific role to ChatGPT as it is a great way to provide context before prompting the language model. We primed the model to be an expert in music and the emotions conveyed in it.

Table 2. Response provided for the Direct approach for the song 'Hello' by Adele.

Direct: ChatGPT Response for 'Hello' by Adele

{
"valence": "0.4",
"arousal": "0.6",
"reasons": ["The valence estimate of 0.4 is based on the melancholic and reflective nature of the song's lyrics, which evoke feelings of sadness and longing.", "The arousal estimate of 0.6 is based on the powerful and emotionally charged vocal performance by Adele, as well as the dynamic range and intensity of the instrumentation."]
}

We identified errors and potential hallucinations by reviewing responses, cross-referencing them, and re-prompting when necessary. Even after thorough review and re-prompts, we found that ChatGPT could not estimate the arousal and valence for a few songs. Nevertheless, the response rate was still very high.

4.1 Method 1: Direct Estimate

Initially, ChatGPT would not output numeric values, stating that it is not trained for such calculations. However, by clarifying the request as a "subjective estimate", ChatGPT then provided numeric values ranging from 0 to 1. This change encourages the LLM to interpret the emotional content of the songs with more flexibility by taking various factors into consideration such as lyrics, reviews, social media, and articles. Although ChatGPT's reasoning process is obscured by the model's algorithm, we hypothesize that subjective estimates may align more closely with the human annotation process since, like human annotators, the model is not calculating exact values but instead drawing on broader contextual cues. An example response soliciting a direct estimate is given in Table 2.

4.2 Method 2: Descriptive Words

In our second approach, we requested that ChatGPT describe each song with five descriptive words. We chose five words after testing ranges from 2-10 words. We observed that requesting fewer than five words led to insufficient information and more than five led to redundancy. We referenced each word in the response from ChatGPT against the corresponding arousal and valence scores using the extended Affective Norms for English Words (XANEW), which comprises 13,915 English lemmas [31], building upon the original list [4]. If a word was not listed, we used the `NLTK` package[4] to find a synonym for the word and matched it to the arousal and valence values. In an approach similar to other studies [18], we took the average of these scores to create a composite arousal and valence score for each song. These values were then normalized to [0-1] to maintain consistency with the direct estimates. An example response is given in Table 3.

[4] https://www.nltk.org/ (last accessed 2026/01/20)

Table 3. Response provided for the 5-word approach for the song 'Hello' by Adele.

5-words: ChatGPT Response for 'Hello' by Adele

{
"words": ["nostalgic", "heartbreak", "longing", "regret", "resignation"],
"reasons": ["The lyrics and Adele's vocal delivery evoke a sense of longing for the past and reminiscing about a lost love.", "The song is about a breakup and the emotions of heartbreak are evident in the lyrics and Adele's powerful vocals.", "The lyrics express a deep yearning for connection and a desire to reconcile with a past lover.", "There is a sense of regret and reflection on past mistakes in the lyrics, which contributes to the emotional content of the song.", "The overall tone of the song conveys a feeling of resignation and acceptance of the end of a relationship."]
}

Table 4. Response for the long-form approach for the song 'Hello' by Adele.

Long-form: ChatGPT Response for 'Hello' by Adele

{
"paragraph": "The song 'Hello' by Adele conveys a deep sense of longing, heartbreak, and nostalgia. Adele's powerful and emotive vocals, combined with the poignant lyrics, evoke a strong emotional response from the listener. The lyrics speak of a past relationship and the desire to reconnect, highlighting the pain of lost love and the yearning for closure. The melancholic melody and soulful instrumentation further enhance the emotional intensity of the song.",
"reasons": "The lyrics of the song express themes of longing, heartbreak, and nostalgia. Adele's vocals and delivery are known for their emotional depth, which adds to the overall sentiment conveyed in the song. The melody and instrumentation also contribute to the emotional intensity felt by the listener."
}

4.3 Method 3: Long-form Description

Next, we tasked ChatGPT to provide long-form textual descriptions of the songs that summarize their emotional content. These descriptions were no longer than 1024 tokens. From the response, we cleaned up the text and removed stop words. For the remaining words in the paragraph, we used a methodology similar to the second approach. Each word, or its synonym, was referenced against the XANEW list to find its corresponding arousal and valence scores. We then averaged the scores of all words or synonyms to obtain the arousal and valence values for each song. These values were also normalized to [0-1] to ensure consistency across all methodologies. An example response is given in Table 4.

5 Statistical Analysis

We assessed ChatGPT's emotion estimation in two ways. First, we compared the results of the three methods (direct, 5-word, long-form). Second, we compared the estimated values to published human and algorithmic annotations. For both comparisons, we report Pearson's correlation coefficient (PCC). We also consider a discretization of the contiguous arousal and valence values into the four quadrants of the circumplex scale, namely low arousal/low valence, low arousal/high valence, high arousal/low valence, and high arousal/high valence. Through classification, we are able to see the percentage of arousal and valence values that are in the same quadrant. This way, if one set of arousal and valence values differ from

Table 5. Comparison of the three approaches across five datasets: 5-words vs. Long-form (left), 5-words vs. Direct (middle), and Long-form vs. Direct (right). Results show PCC for arousal and valence, and accuracy for quadrant classification.

	5-words vs. Longform			**5-words vs. Direct**			**Longform vs. Direct**		
	Aro.	Val.	Quad.	Aro.	Val.	Quad.	Aro.	Val.	Quad.
AMG1608	0.685	0.772	59.3%	0.632	0.726	61.0%	0.600	0.700	54.8%
DEAM	0.571	0.623	55.1%	0.446	0.629	54.5%	0.447	0.610	47.0%
DEAP	0.712	0.843	69.2%	0.626	0.822	56.9%	0.541	0.796	56.9%
Deezer	0.626	0.707	54.8%	0.512	0.675	48.5%	0.459	0.684	44.7%
PMEmo	0.501	0.716	52.5%	0.410	0.621	54.7%	0.357	0.631	40.3%

Table 6. The mean ± std of arousal across the 5 different datasets and 4 methods.

Dataset	**Direct**	**5-Words**	**Longform**	**Original**
AMG1608	0.63 ± 0.15	0.57 ± 0.20	0.52 ± 0.18	0.55 ± 0.22
DEAM	0.60 ± 0.15	0.57 ± 0.19	0.49 ± 0.17	0.49 ± 0.20
DEAP	0.65 ± 0.16	0.50 ± 0.29	0.56 ± 0.25	0.49 ± 0.21
Deezer	0.61 ± 0.14	0.48 ± 0.19	0.47 ± 0.15	0.50 ± 0.19
PMEmo	0.67 ± 0.14	0.52 ± 0.16	0.45 ± 0.15	0.62 ± 0.18

another set, we can still assess the percentage of songs that reside in the same quadrant. This enables us to gain insights into high-level categorical similarities between emotions despite any numerical differences.

Lastly, we examined the five existing datasets to find any overlapping songs in either two, three, four, or all five of these datasets. For these overlapping songs, we calculated the PCC to evaluate the level of agreement between the original published annotations. Through this analysis, we were able to determine how well the original annotations align with each other. This provided insights into the consistency and reliability of emotion recognition across different datasets.

5.1 Comparison of Labeling Approaches

First, we compare the similarity of our annotations produced by the three methods of annotation. These results are shown in Table 5. When comparing these methods, we observe that the 5-words and long-form extraction approaches share a higher correlation for both valence and arousal when compared to direct/5-words and to direct/long-form approaches. For quadrant classification, we observe a similar trend; the classification accuracy between the 5-words and long-form approach is generally higher when either is compared to the direct estimations. Overall, these results show that the 5-word approach and long-form method yield more similar results compared to the direct approach; this could be explained by the similar text-processing approach in the two methods. We also observe that the correlation for valence tends to be higher than that for arousal across all three comparisons in each of the five datasets. This likely indicates that valence, which measures negative to positive, is much easier to interpret than arousal, which measures calm to exciting. Diving deeper into specific datasets, we find that

Table 7. The mean ± std of valence across the 5 different datasets and 4 methods.

Dataset	Direct	5-Words	Longform	Original
AMG1608	0.56 ± 0.20	0.62 ± 0.23	0.50 ± 0.18	0.51 ± 0.17
DEAM	0.57 ± 0.18	0.69 ± 0.23	0.54 ± 0.18	0.49 ± 0.17
DEAP	0.54 ± 0.23	0.52 ± 0.30	0.50 ± 0.23	0.49 ± 0.25
Deezer	0.52 ± 0.18	0.63 ± 0.20	0.51 ± 0.16	0.56 ± 0.29
PMEmo	0.62 ± 0.18	0.66 ± 0.22	0.60 ± 0.18	0.60 ± 0.16

Table 8. Comparison of the original and ChatGPT annotations for Direct (left), 5-words (middle), and Long-form (right) approaches across five datasets. Results show PCC scores for arousal and valence, and accuracy for quadrant classification.

	Direct			**5-words**			**Longform**		
	Aro.	Val.	Quad.	Aro.	Val.	Quad.	Aro.	Val.	Quad.
AMG1608	0.582	0.635	53.6%	0.534	0.597	52.8%	0.532	0.579	54.7%
DEAM	0.185	0.201	32.5%	0.263	0.160	36.0%	0.240	0.101	35.1%
DEAP	0.560	0.688	49.1%	0.580	0.637	60.8%	0.496	0.637	55.0%
Deezer	0.208	0.446	35.6%	0.274	0.398	38.9%	0.284	0.394	41.2%
PMEmo	-0.091	0.044	44.0%	0.016	0.033	37.1%	-0.002	0.004	25.5%

DEAP consistently has higher correlation across all three comparisons whereas, PMEmo and DEAM tend to have comparatively lower correlations. This trend may indicate that the emotions expressed in the musical pieces in the DEAP dataset are much more clearly defined, which leads to more consistent correlations across all the three methods tested. For each dataset and method, we provide the mean and standard deviation for arousal as Table 6 and for valence as Table 7.

5.2 Comparison to Existing Annotations

Next, we compare the performance of our three LLM-based approaches with the human- (AMG1608, DEAM, DEAP, PMEmo) and algorithmically-annotated (Deezer) data. The goal of this comparison is to investigate the similarity of our LLM-derived annotations with the existing annotations widely used in MER research. The results of this statistical comparison are given in Table 8.

We find that the LLM-generated annotations for the DEAP dataset generally achieved higher correlations to human annotations, followed by AMG1608. This may reflect the fact that both datasets were created using rigorous experimental designs employing many different human annotators. PMEmo, on the other hand, yielded the lowest agreement, perhaps explained by the less rigorous experimental design choices in that study. Given that most of the annotators were non-English speakers, this could have made it harder for our LLM-based approaches to align well with the original annotations. DEAM has the second lowest correlation values across all three approaches, which could be explained by its use of royalty-free music, which is typically less popular and it is possible that ChatGPT was not extensively trained on these songs. In this comparison, we also observe that correlations for valence are generally higher than arousal, which again suggests

Table 9. PCC of original annotations for the songs (n=294) that appear in more than one dataset. Correlations for arousal are given to the left of the diagonal and correlations for valence are given right of the diagonal. The parentheses contain the total count of the overlapping songs between any two datasets.

	AMG	**DEAM**	**DEAP**	**Deezer**	**PmEmo**
AMG	\	— (0)	-0.353 (3)	0.639 (246)	-0.254 (4)
DEAM	— (0)	\	— (0)	0.576 (3)	— (0)
DEAP	0.901 (3)	— (0)	\	0.705 (29)	— (0)
Deezer	0.571 (246)	-0.374 (3)	0.484 (29)	\	-0.014 (9)
PMEmo	0.257 (4)	— (0)	— (0)	-0.001 (9)	\

that the LLMs might find it easier to estimate valence than to estimate arousal. One noteworthy point is that, while the correlation values for PMEmo are quite low, the quadrant classifications are relatively higher, indicating some success in the model's ability to estimate the general emotion. Overall, we found modest agreement between our LLM estimations and the original data, but this agreement varies significantly across the different datasets with trends that are similar to those observed in Section 5.1.

5.3 Comparison between Human Annotations

In human annotation studies, inter-rater reliability is often poor [27]. Because the extant MER datasets used different methodologies, we questioned the agreement between raters across these studies. We investigated the correlation between human annotations of the same song produced through different experimental designs. We examined the five existing datasets to find any songs that exist in more than one dataset. However, we found only 294 songs that overlapped in more than one datasets. The counts and correlations are shown in Table 9.

When examining the consistency of annotations across different datasets for the same song, we found the correlation values vary significantly. This likely results from differences in the experimental approach and underscores the subjective nature of the annotation task. This observation emphasizes the need for a consistent and scalable approach to annotate MER datasets. Although only 294 songs overlapped across the datasets, our LLM-derived annotations show higher correlations to the original annotations and between the three approaches. This suggests potential to achieve more consistency with LLM based annotations.

6 Discussion

We present a novel approach to address challenges faced when creating annotated datasets for MER. We use an LLM to estimate arousal and valence using a direct numeric approach, a 5-word indirect estimation, and a long-form answer indirect estimation. This approach has potential to facilitate the generation of synthetic annotations on a large scale in a systematic manner. Our findings indicate that LLMs can contribute valuable insights into the emotional content of music.

6.1 Limitations and Future Work

Given the novelty of our approach, we acknowledge certain limitations which motivate our future work. One important concern is the tendency for LLMs to occasionally 'hallucinate'. To avoid this, we took steps to craft specific prompts and to identify hallucinations by manually inspecting the output and re-prompting the language models when appropriate. However, it is difficult to completely eliminate instances of hallucination. One solution to reduce hallucination could be to refine our prompt designs to include strategies such as Chain-of-Thought verification [20]. Another solution could necessitate re-prompting each song multiple times and choosing the most common response. However, this approach would limit scalability and increase financial costs. Another promising direction would be to explore how other language models perform with annotation estimation. While this study focuses on ChatGPT, we could compare other LLM models such as LLaMA or Claude. This comparison would be very beneficial to find an ensemble of models best suited for annotating music.

In this study, we focused on songs in existing datasets to facilitate comparison with human annotations. As future work, we plan to create a larger and genre-diverse dataset. We believe such a dataset is crucial for developing accurate, real-world emotion recognition systems, even if the annotations are indirect and imperfect estimations. Furthermore, incorporating other emotional classification scales, such as Plutchik's wheel of emotions [24], will provide researchers other avenues to explore the link between music and emotion.

6.2 Conclusion

This study presents a novel approach that leverages a LLM to estimate an average emotive reaction to pieces of music. This approach demonstrates the potential to generate emotion annotations which, although synthetic, have a moderate and positive correlation with emotions reported by humans. We emphasize that this relationship is comparable, if not stronger, to the level of correlation seen between humans annotating the same songs but in different settings. Ultimately, the true effectiveness of these annotations in practical applications, such as music recommendation systems [21] or playlist generation [9], is yet to be determined. Nevertheless, MER desperately needs large reliable datasets and this research invites further exploration of LLMs to estimate human reactions to music.

References

1. Aljanaki, A., Yang, Y.H., Soleymani, M.: Developing a benchmark for emotional analysis of music. PloS one **12**(3), 1–22 (2017). https://doi.org/10.1371/journal.pone.0173392
2. Beery, A., Donnelly, P.J.: Learning affective responses to music from social media discourse. In: Abbas, M. (ed.) Practical Solutions for Diverse Real-World NLP Applications, pp. 93–119. Springer International Publishing, Cham (2024). https://doi.org/10.1007/978-3-031-44260-5_6
3. Bertin-Mahieux, T., Ellis, D.P.W., Whitman, B., Lamere, P.: The million song dataset. In: Klapuri, A., Leider, C. (eds.) Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR). pp. 591–596. ISMIR, Miami, USA (Oct 2011). https://doi.org/10.5281/zenodo.1415820
4. Bradley, M.M., Lang, P.J.: Affective norms for English words (ANEW): Instruction manual and affective ratings. Tech. rep., Technical report C-1, the Center for Research in Psychophysiology (1999), https://api.semanticscholar.org/CorpusID:145474983, last accessed 2025/11/10
5. Çano, E., Morisio, M.: Moodylyrics: A sentiment annotated lyrics dataset. In: Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence (ISMSI). p. 118–124. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3059336.3059340
6. Chen, Y.A., Yang, Y.H., Wang, J.C., Chen, H.: The AMG1608 dataset for music emotion recognition. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). p. 693–697. IEEE (Apr 2015). https://doi.org/10.1109/ICASSP.2015.7178058
7. Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J., Moussallam, M.: Music mood detection based on audio and lyrics with deep neural net. In: Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR). pp. 370–375. ISMIR (2018). https://doi.org/10.5281/zenodo.1492427
8. Ekman, P.: Expression and the nature of emotion. In: Scherer, K., Ekman, P. (eds.) Approaches to Emotion, vol. 3:19, pp. 319–344. Lawrence Erlbaum, Hillsdale, NJ (1984). https://doi.org/10.4324/9781315798806
9. Gaur, S., Donnelly, P.J.: Generating smooth mood-dynamic playlists with audio features and KNN. In: Johnson, C., Rebelo, S.M., Santos, I. (eds.) Proceedings of the 13th International Conference of Evolutionary and Biologically Inspired Music, Sound, Art and Design (EvoMUSART). vol. 14633, pp. 162–178. Springer (2024). https://doi.org/10.1007/978-3-031-56992-0_11
10. Gómez-Cañón, J.S., Cano, E., Eerola, T., Herrera, P., Hu, X., Yang, Y.H., Gómez, E.: Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications. IEEE Signal Processing Magazine **38**(6), 106–114 (2021). https://doi.org/10.1109/MSP.2021.3106232
11. Gómez-Cañón, J.S., Gutiérrez-Páez, N., Porcaro, L., Porter, A., Cano, E., Herrera-Boyer, P., Gkiokas, A., Santos, P., Hernández-Leo, D., Karreman, C., et al.: TROMPA-MER: an open dataset for personalized music emotion recognition. Journal of Intelligent Information Systems **60**(2), 549–570 (2023). https://doi.org/10.1007/s10844-022-00746-0
12. Hevner, K.: Experimental studies of the elements of expression in music. The American Journal of Psychology **48**(2), 246–268 (1936)
13. Hu, X., Downie, J.S., Ehmann, A.F.: Lyric text mining in music mood classification. In: Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR). pp. 411–416 (2009). https://doi.org/10.5281/zenodo.1416790

14. Hult, S., Kreiberg, L.B., Brandt, S.S., Jónsson, B.T.: Analysis of the effect of dataset construction methodology on transferability of music emotion recognition models. In: Proceedings of the 2020 International Conference on Multimedia Retrieval. pp. 316–320 (2020). https://doi.org/10.1145/3372278.3390733
15. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12), 1–38 (2023). https://doi.org/10.1145/3571730
16. Khanal, P., Donnelly, P.J.: Emotionotes dataset: Decoding emotions in classical music through concert program notes. In: International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar). pp. 339–356. Springer (2025). https://doi.org/0.1007/978-3-031-90167-6_23
17. Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Patras, I.: DEAP: Database for emotion analysis using physiological signals. IEEE transactions on affective computing **3**(1), 18–31 (2011). https://doi.org/10.1109/T-AFFC.2011.15
18. Laurier, C., Sordo, M., Serra, J., Herrera, P.: Music mood representations from social tags. In: Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR). pp. 381–386 (2009)
19. Liebetrau, J., Schneider, S.: Music and emotions: a comparison of measurement methods. In: Audio Engineering Society Convention 134. Audio Engineering Society (2013), https://publica.fraunhofer.de/handle/publica/381224, last accessed 2026/01/30
20. Ling, Z., Fang, Y., Li, X., Huang, Z., Lee, M., Memisevic, R., Su, H.: Deductive verification of chain-of-thought reasoning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, vol. 36, pp. 36407–36433. Curran Associates, Inc., Red Hook, NY, USA (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/72393bd47a35f5b3bee4c609e7bba733-Paper-Conference.pdf, last accessed 2026/01/30
21. Liu, Z., Xu, W., Zhang, W., Jiang, Q.: An emotion-based personalized music recommendation framework for emotion improvement. Information Processing & Management **60**(3), 103256 (2023). https://doi.org/10.1016/j.ipm.2022.103256
22. Panda, R., Malheiro, R., Paiva, R.P.: Novel audio features for music emotion recognition. IEEE Transactions on Affective Computing **11**(4), 614–626 (2020). https://doi.org/10.1109/TAFFC.2018.2820691
23. Panda, R., Malheiro, R., Paiva, R.P.: Audio features for music emotion recognition: A survey. IEEE Transactions on Affective Computing **14**(1), 68–88 (2023). https://doi.org/10.1109/TAFFC.2020.3032373
24. Plutchik, R.: A general psychoevolutionary theory of emotion. In: Plutchik, R., Kellerman, H. (eds.) Theories of Emotion, pp. 3–33. Academic Press (1980). https://doi.org/10.1016/B978-0-12-558701-3.50007-7
25. Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology **39**(6), 1161–1178 (Dec 1980). https://doi.org/10.1037/h0077714
26. Russell, J.A., Weiss, A., Mendelsohn, G.A.: Affect grid: a single-item scale of pleasure and arousal. Journal of Personality and Social Psychology **57**(3), 493 (1989). https://doi.org/10.1037/0022-3514.57.3.493
27. Schedl, M., Gómez, E., Trent, E.S., Tkalčič, M., Eghbal-Zadeh, H., Martorell, A.: On the interrelation between listener characteristics and the perception of emotions in classical orchestra music. IEEE Transactions on Affective Computing **9**(4), 507–525 (2018). https://doi.org/10.1109/TAFFC.2017.2663421

28. Soleymani, M., Kierkels, J.J., Chanel, G., Pun, T.: A Bayesian framework for video affective representation. In: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. pp. 1–7. IEEE (2009). https://doi.org/10.1109/ACII.2009.5349563
29. Sylolypavan, A., Sleeman, D., Wu, H., Sim, M.: The impact of inconsistent human annotations on ai driven clinical decision making. NPJ Digital Medicine **6**(1), 26 (2023). https://doi.org/10.1038/s41746-023-00773-3
30. Wang, J.C., Yang, Y.H., Chang, K., Wang, H.M., Jeng, S.K.: Exploring the relationship between categorical and dimensional emotion semantics of music. In: Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies. pp. 63–68 (2012). https://doi.org/10.1145/2390848.2390865
31. Warriner, A.B., Kuperman, V., Brysbaert, M.: Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior Research Methods **45**, 1191–1207 (2013). https://doi.org/10.3758/s13428-012-0314-x
32. Watcharasupat, K.N., Ding, Y., Ma, T.A., Seshadri, P., Lerch, A.: Uncertainty estimation in the real world: A study on music emotion recognition. In: European Conference on Information Retrieval. pp. 218–232. Springer (2025). https://doi.org/0.1007/978-3-031-88711-6_14
33. Zhang, K., Zhang, H., Li, S., Yang, C., Sun, L.: The PMEmo dataset for music emotion recognition. In: Proceedings of the 2018 International Conference on Multimedia Retrieval. p. 135–142. ACM, Yokohama Japan (Jun 2018). https://doi.org/10.1145/3206025.3206037
34. Zhang, Y., Ding, X., Gu, N.: Understanding fatigue and its impact in crowdsourcing. In: 22nd International Conference on Computer Supported Cooperative Work in Design (CSCWD). pp. 57–62. IEEE (2018). https://doi.org/10.1109/CSCWD.2018.8465305
35. Zhuo, L., Yuan, R., Pan, J., Ma, Y., LI, Y., Zhang, G., Liu, S., Dannenberg, R., Fu, J., Lin, C., Benetos, E., Chen, W., Xue, W., Guo, Y.: Lyricwhiz: Robust multilingual zero-shot lyrics transcription by whispering to ChatGPT. Proceedings of the 24th International Society for Music Information Retrieval Conference p. 343–351 (2023). https://doi.org/10.5281/zenodo.10265295