

Oregon State University

Addressing Dataset Scarcity in Music Emotion Recognition with LLMs

EvoMUSART 2026

Janita Aamir and Patrick J. Donnelly

April 9, 2026

soundbendor lab

Music Emotion Recognition (MER)

Task

To automatically identify emotions elicited by pieces of music

Music Emotion Recognition (MER)

Task

To automatically identify emotions elicited by pieces of music

Approaches:

- **audio signals** [1]
- acoustic features [2]
- physiological signals [3]
- song lyrics [4]
- social media discourse [5]
- concert program notes [6]

Music Emotion Recognition (MER)

Task

To automatically identify emotions elicited by pieces of music

Approaches:

- **audio signals** [1]
- acoustic features [2]
- physiological signals [3]
- song lyrics [4]
- social media discourse [5]
- concert program notes [6]

Applications:

- music recommendation [7]
- mood-based playlists [8]
- music therapy [9]
- musicology research
- music captioning
- adaptive soundtracks

Motivation

Problem

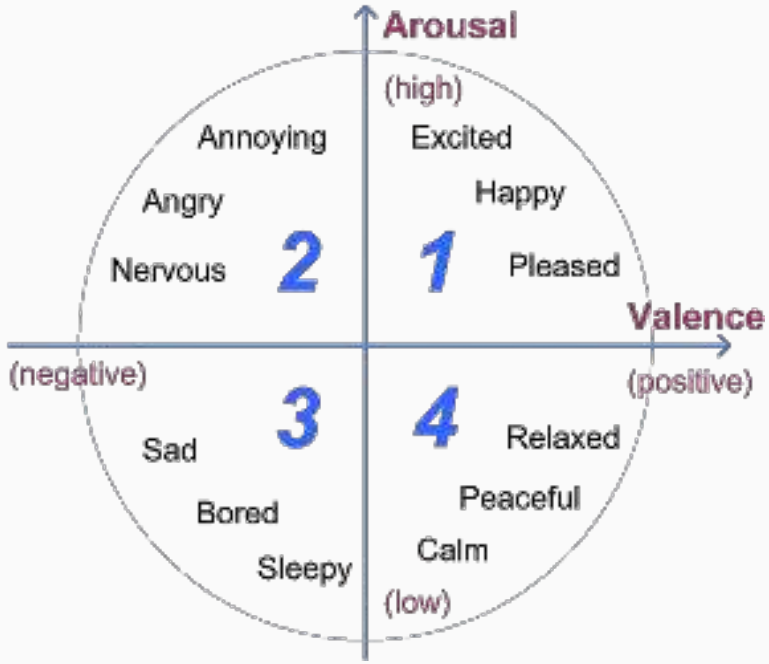
There is a lack of large, diverse datasets for MER

Annotating datasets for music emotion is:

- expensive
- time-consuming
- difficult to define
- subjective [10]
- unreliable [11]

Circumplex model for emotions

- Russell's circumplex model [12]
  - emotions as continuous values
- **Valence**
  - unpleasant to pleasant
- **Arousal**
  - calm to excited



Circumplex model for emotions

Datasets

Summary of Datasets

Dataset		Count	Method
AMG1608	[13]	1608	Online
DEAM	[14]	1820	Online
DEAP	[3]	120	Lab
Deezer	[15]	18,644	Algorithmic
PMEmo	[16]	794	Lab

AMG1608 Dataset

All **M**usic **G**uide of **1608** songs (AMG1608)

- Chen, Yang, Wang, *et al.* (2015) [13]

Strengths:

- 1608 songs
- 655 different raters
- multiple raters per song
- each rater rated 13 songs

Weaknesses:

- short 30-second excerpts
- popular Western music genres
- crowd-sourced online
- risk of annotation fatigue

DEAM Dataset

MediaEval **D**atabase for **E**motional **A**nalysis in **M**usic (DEAM)

- Aljanaki, Yi-Hsuan, & Soleymani (2017) [14]

Strengths:

- 1802 songs
- multiple raters per song
- dynamic annotations over time
- audio available

Weaknesses:

- 45-second excerpts
- popular Western music genres
- crowd-sourced online
- obscure royalty-free songs

DEAP Dataset

Database for **E**motion **A**nalysis using **P**hysiological Signals (DEAP)

- Koelstra, Muhl, Soleymani, *et al.* (2011) [14]

Strengths:

- multimodal (video & audio)
- collected physiological signals
- addressed annotation fatigue
- balanced across quadrants

Weaknesses:

- 120 songs
- 60-second excerpts
- biased for "emotional" songs
- coupled video and audio

Deezer Dataset

Deezer mood detection dataset (Deezer)

- Delbouys, Hennequin, Piccoli, *et al.* (2018) [15]

Strengths:

- 18,644 songs
- novel approach for scale
- mapping to MSD

Weaknesses:

- algorithmically-generated
- only 2,762 unique values
- no audio or lyrics provided

PMEmo Dataset

Physiological–**M**usic **Emo**tions (PMEmo)

- Zhang, Zhang, Li, *et al.* (2018) [16]

Strengths:

- 794 songs
- 457 different raters
- multiple raters per song
- lab-based study
- audio released (but removed)

Weaknesses:

- only trending pop songs
- Chinese subjects listening to English lyrics
- copyright-concerns
- biased collection concerns

Correlations between Human 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)	\

Correlations of Arousal (left of diagonal) and Valence (right)

Approach

Methods

Approach

Task `ChatGPT` to estimate the valence and arousal ratings of music

Methods

Approach

Task `ChatGPT` to estimate the valence and arousal ratings of music

ChatGPT has access to:

- song lyrics
- analyses
- press reviews
- online discussions
- interviews

Methods

Approach

Task `ChatGPT` to estimate the valence and arousal ratings of music

ChatGPT has access to:

- song lyrics
- analyses
- press reviews
- online discussions
- interviews

Structured prompts:

- reduce hallucinations [17]
- clear task descriptions
- require reasoning
- JSON output format
- re-prompting when necessary

Prompting Methods

Three Prompts to ChatGPT:

1. direct, numeric estimate
2. list of five emotion words
3. long-form textual description

Estimating emotions from words:

- **Reference** dataset [18] of valence and arousal scores
- **Lookup** scores of individual emotion words
- **Average** the word ratings together

Example Prompt for 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.

Example Response: Direct

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."]
}

Response provided for the Direct approach for the song 'Hello' by Adele

Example Response: 5-word

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."]
}

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

Example Response: Long-form

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."
}

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

Results

Comparison of ChatGPT Approaches

	5-words vs. Longform		5-words vs. Direct		Longform vs. Direct	
	Aro.	Val.	Aro.	Val.	Aro.	Val.
AMG1608	0.685	0.772	0.632	0.726	0.600	0.700
DEAM	0.571	0.623	0.446	0.629	0.447	0.610
DEAP	0.712	0.843	0.626	0.822	0.541	0.796
Deezer	0.626	0.707	0.512	0.675	0.459	0.684
PMEmo	0.501	0.716	0.410	0.621	0.357	0.631

Pearson's correlation for Arousal and Valence comparing ChatGPT annotations

Arousal and Valence Distributions

	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

Mean ± std values for Arousal

	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

Mean ± std values for Valence

Comparison of ChatGPT vs. Human

	Direct		5-words		Longform	
	Aro.	Val.	Aro.	Val.	Aro.	Val.
AMG1608	0.582	0.635	0.534	0.597	0.532	0.579
DEAM	0.185	0.201	0.263	0.160	0.240	0.101
DEAP	0.560	0.688	0.580	0.637	0.496	0.637
Deezer	0.208	0.446	0.274	0.398	0.284	0.394
PMEmo	-0.091	0.044	0.016	0.033	-0.002	0.004

Pearson's correlation for Arousal and Valence for human vs. ChatGPT

Summary

Contributions

- Novel approach to estimate emotion using LLMs
  - direct numeric values
  - five key words
  - long-form descriptions
- New dataset of emotion pseudo-annotations
- Comparison of existing human-labeled datasets
  - showed that human annotations differ across datasets
- Enabling MER annotations at large scale

Limitations

- Indirect approach results in pseudo-annotations
  - reasoning is obscured by the LLM process
  - LLMs may produce hallucinations
  - different LLMs may produce different results
- Word descriptions may not match audio characteristics
- API costs associated with LLM procedures
- Data lacks diversity of music genres, cultures

Future Work

- Consult an ensemble of multiple LLMs
  - ChatGPT, Llama, Gemini, Claude, Mistral
- Collect more data
  - Billboard Top 100, Apple Music Top 100
  - explore less common genres, world music
- Estimate emotion directly from the music
  - audio signals
  - acoustic features
  - spectrograms

References i

[1] N. He and S. Ferguson, **"Multi-view neural networks for raw audio-based music emotion recognition,"** in *2020 IEEE International Symposium on Multimedia (ISM)*, 2020, pp. 168–172. DOI: 10.1109/ISM.2020.00037.

[2] R. Panda, R. Malheiro, and R. P. Paiva, **"Audio features for music emotion recognition: A survey,"** *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 68–88, 2023. DOI: 10.1109/TAFFC.2020.3032373.

[3] S. Koelstra et al., **"DEAP: Database for emotion analysis using physiological signals,"** *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011. DOI: 10.1109/T-AFFC.2011.15.

[4] E. Çano and M. Morisio, **"Moodylyrics: A sentiment annotated lyrics dataset,"** in *Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence (ISMSI)*, Hong Kong, Hong Kong: Association for Computing Machinery, 2017, pp. 118–124, ISBN: 9781450347983. DOI: 10.1145/3059336.3059340.

[5] A. Beery and P. J. Donnelly, **"Learning affective responses to music from social media discourse,"** in *Practical Solutions for Diverse Real-World NLP Applications*, M. Abbas, Ed., Cham: Springer International Publishing, 2024, pp. 93–119, ISBN: 978-3-031-44260-5. DOI: 10.1007/978-3-031-44260-5_6.

[6] P. Khanal and P. J. Donnelly, **"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)*, Springer, 2025, pp. 339–356. DOI: 0.1007/978-3-031-90167-6_23.

[7] Z. Liu, W. Xu, W. Zhang, and Q. Jiang, **"An emotion-based personalized music recommendation framework for emotion improvement,"** *Information Processing & Management*, vol. 60, no. 3, p. 103 256, 2023, ISSN: 0306-4573. DOI: 10.1016/j.ipm.2022.103256.

[8] S. Gaur and P. J. Donnelly, **"Generating smooth mood-dynamic playlists with audio features and KNN,"** in *Proceedings of the 13th International Conference of Evolutionary and Biologically Inspired Music, Sound, Art and Design (EvoMUSART)*, C. Johnson, S. M. Rebelo, and I. Santos, Eds., vol. 14633, Springer, 2024, pp. 162–178, ISBN: 978-3-031-56992-0. DOI: 10.1007/978-3-031-56992-0_11.

[9] H. Feng, M. H. Mahoor, and F. Dino, **"A music-therapy robotic platform for children with autism: A pilot study,"** *Frontiers in Robotics and AI*, vol. 9, 2022, ISSN: 2296-9144. DOI: 10.3389/frobt.2022.855819.

[10] M. Schedl, E. Gómez, E. S. Trent, M. Tkalčič, H. Eghbal-Zadeh, and A. Martorell, **"On the interrelation between listener characteristics and the perception of emotions in classical orchestra music,"** *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 507–525, 2018. DOI: 10.1109/TAFFC.2017.2663421.

References ii

[11] Y. Zhang, X. Ding, and N. Gu, **"Understanding fatigue and its impact in crowdsourcing,"** in *22nd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, IEEE, 2018, pp. 57–62. DOI: 10.1109/CSCWD.2018.8465305.

[12] R. Plutchik, **"A general psychoevolutionary theory of emotion,"** in *Theories of Emotion*, R. Plutchik and H. Kellerman, Eds., Academic Press, 1980, pp. 3–33, ISBN: 978-0-12-558701-3. DOI: 10.1016/B978-0-12-558701-3.50007-7.

[13] Y.-A. Chen, Y.-H. Yang, J.-C. Wang, and H. Chen, **"The AMG1608 dataset for music emotion recognition,"** en, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Apr. 2015, pp. 693–697, ISBN: 978-1-4673-6997-8. DOI: 10.1109/ICASSP.2015.7178058.

[14] A. Aljanaki, Y.-H. Yang, and M. Soleymani, **"Developing a benchmark for emotional analysis of music,"** *PloS one*, vol. 12, no. 3, pp. 1–22, 2017. DOI: 10.1371/journal.pone.0173392.

[15] R. Delbouys, R. Hennequin, F. Piccoli, J. Royo-Letelier, and M. Moussallam, **"Music mood detection based on audio and lyrics with deep neural net,"** en, in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, ISMIR, 2018, pp. 370–375. DOI: 10.5281/zenodo.1492427.

[16] K. Zhang, H. Zhang, S. Li, C. Yang, and L. Sun, **"The PMEmo dataset for music emotion recognition,"** en, in *Proceedings of the 2018 International Conference on Multimedia Retrieval*, Yokohama Japan: ACM, Jun. 2018, pp. 135–142, ISBN: 978-1-4503-5046-4. DOI: 10.1145/3206025.3206037.

[17] L. Zhuo et al., **"Lyricwhiz: Robust multilingual zero-shot lyrics transcription by whispering to ChatGPT,"** *Proceedings of the 24th International Society for Music Information Retrieval Conference*, pp. 343–351, 2023. DOI: 10.5281/zenodo.10265295.

[18] A. B. Warriner, V. Kuperman, and M. Brysbaert, **"Norms of valence, arousal, and dominance for 13,915 English lemmas,"** *Behavior Research Methods*, vol. 45, pp. 1191–1207, 2013. DOI: 10.3758/s13428-012-0314-x.

Thank You!