GLOBALMOOD: A CROSS-CULTURAL BENCHMARK FOR MUSIC EMOTION RECOGNITION

Harin Lee[1,2,3] **Elif Çelen**[1] **Peter Harrison**[4] **Manuel Anglada-Tort**[5]
Pol van Rijn[1] **Minsu Park**[6] **Marc Schönwiesner**[3] **Nori Jacoby**[1,7]
[1]Max Planck Institute for Empirical Aesthetics
[2]Max Planck Institute for Human Cognitive and Brain Sciences
[3]Leipzig University [4]University of Cambridge [5]Goldsmiths, University of London
[6]New York University Abu Dhabi [7]Cornell University
harin.lee@ae.mpg.de

ABSTRACT

Human annotations of mood in music are essential for music generation and recommender systems. However, existing datasets predominantly focus on Western songs with mood terms derived from English, which may limit generalizability across diverse linguistic and cultural backgrounds. To address this, we introduce 'GlobalMood', a novel cross-cultural benchmark dataset comprising 1,180 songs sampled from 59 countries, with large-scale annotations collected from 2,519 individuals across five culturally and linguistically distinct locations: U.S., France, Mexico, S. Korea, and Egypt. Rather than imposing predefined mood categories, we implement a bottom-up, participant-driven approach to organically elicit culturally specific music-related mood terms. We then recruit another pool of human participants to collect 988,925 ratings for these culture-specific descriptors. Our analysis confirms the presence of a valence-arousal structure shared across cultures, yet also reveals significant divergences in how certain mood terms, despite being dictionary equivalents, are perceived cross-culturally. State-of-the-art multimodal models benefit substantially from fine-tuning on our cross-culturally balanced dataset, as evidenced by improved alignment with human evaluations—particularly in non-English contexts. More broadly, our findings inform the ongoing debate on the universality versus cultural specificity of emotional descriptors, and our methodology can contribute to other multimodal and cross-lingual research.

1. INTRODUCTION

Music evokes diverse emotional responses in listeners, spanning a wide spectrum beyond basic emotional categories [1, 2]. A central challenge in Music Information Retrieval (MIR) is designing algorithms that can replicate this emotional sensitivity. This is crucial for building recommendation systems that align with listeners' mood and context [3,4], and for generating music that resonates with individual preferences [5]. More broadly, understanding how music conveys emotion is a core question in the science of music [6, 7]. To date, however, most algorithms have been trained on datasets derived from Western listeners and Western music, using mood taxonomies primarily based on English language (e.g., MIREX [8]).

A significant challenge is creating cross-cultural models capable of handling non-Western music and mood vocabularies beyond English. Addressing this challenge is essential to developing algorithms that accurately reflect global users' preferences, including those whose musical tastes extend beyond the limited range of styles currently represented in training datasets. Moreover, without capturing culturally specific mood terms, especially those difficult to translate, key aspects of musical meaning may be missed entirely. Direct dictionary translations of English terms may be insufficient, as mood terms are deeply cultural and may lack exact equivalents [9–12].

To address these issues, we introduce 'GlobalMood', a new benchmark dataset designed to support culturally inclusive and linguistically diverse mood recognition in music. Our contribution innovates along three key dimensions: (i) the diversity of musical stimuli, drawn from 59 countries; (ii) the diversity of annotators, spanning five distinct cultural and linguistic regions; (iii) a data-driven approach for collecting mood descriptors, generated organically by participants in their own language during the annotation process.

Data were collected through two stages involving a total of 2,519 participants and 1,180 songs balanced evenly across 59 countries: In the first stage (Section 4.1; Figure 1), using a smaller subset of 200 songs, we employed a recently developed iterative task that combines open-ended elicitation with collective refinement [11, 13, 14]. Rather than asking listeners to choose from a fixed list of predefined mood terms, we asked them to describe the perceived mood conveyed by a piece of music using free-text tags in their native language, at the same time, rate the tags provided by previous listeners. This approach was key to uncovering mood terms that would otherwise be overlooked by predefined, English-based taxonomies (such as 'appeal/plead' that appears in Korean only).

In the second stage (Section 4.2; Figure 2), we selected the top 20 elicited terms per language and crowdsourced ratings for each tag across the entire set of 1,180 songs.

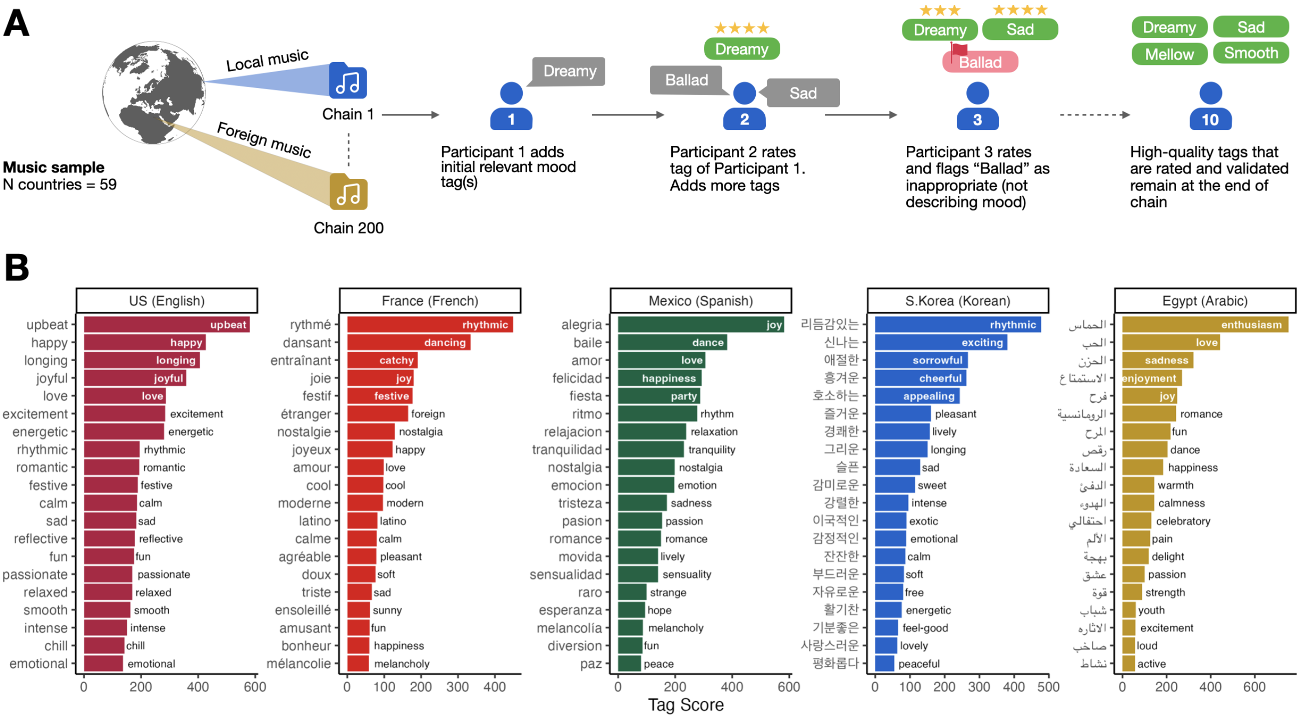

Figure 1. Elicitation and refinement of music mood terms through iterative participant chains. **(A)** Schematic illustration of the collaborative tagging process within a participant chain. Participants contribute new emotion-related word tags for each song, rate the relevance of existing tags, and can also flag irrelevant content, creating a dynamic refinement system. **(B)** The twenty most reliable emotion tags in each language, ranked by their tag scores. Y-axis labels display tags in their original language (left) and English translations (right).

This resulted in a total of 988,925 mood ratings, creating the most comprehensive open-source cross-cultural emotion annotation dataset in Music Emotion Recognition (MER) to date (available at: `https://osf.io/8c5f2`).

We leveraged GlobalMood to test several recent multimodal and multilingual models (Gemini, CLAP) by evaluating their performance on 15-second audio clips under zero-shot, few-shot, and fine-tuned scenarios (Section 4.3; Figure 3). Our results demonstrate that models trained exclusively on English data underperform when applied to certain cultural contexts, but fine-tuning with our cross-cultural dataset significantly improves performance in non-English countries. This highlights the critical importance of cross-cultural data in both training MER models and establishing appropriate benchmarks for their evaluation.

2. RELATED WORKS

2.1 Music Mood Annotation Datasets

Several datasets have been developed for MER systems with varying annotation approaches. Early examples include the widely used MIREX 2007 mood dataset [8] with 240-250 Western songs in five mood clusters derived from AllMusic's English tags (e.g., 'passionate–rousing', 'wistful–bittersweet'), and CAL500 [15] with 500 Western pop/rock songs annotated using 18 English mood terms by U.S. undergraduate listeners. Over time, larger datasets appeared: the DEAM corpus (MediaEval 'Emotion in Music' dataset [16]) containing 2,058 song excerpts with continuous valence/arousal annotations; mood tags mined from large corpora of Spotify music playlists [17]; and the MTG-Jamendo dataset [18], which provides mood/theme tags for 18,486 songs. Notably, Jamendo's tags were freely crowdsourced (56 unique mood labels), which introduced more label variety but still almost entirely in English.

A common limitation across these datasets is their reliance on predefined English descriptors, many of which stem from Western music psychology. For instance, the Geneva Emotional Music Scale (GEMS) defines 45 emotion descriptors (e.g., 'joyful activation') based on studies with European listeners [1], and this taxonomy has been used to annotate datasets like Emotify [19]. Similarly, the mood categories in MIREX and CAL500 were fixed in advance (drawn from AllMusic or prior literature) and presented to annotators as a closed set of options. Consequently, these top-down approaches restrict annotators to the moods the researchers envisioned, leaving any unlisted mood nuances uncaptured and undocumented.

Acknowledging these limitations, recent research has begun exploring MER beyond the Western-centric scope. Hu et al. [20] examined mood annotations of K-pop songs provided by both Korean and American listeners. Their approach involved translating the original MIREX mood categories into Korean for local annotators (included in the 2014 MIREX challenge). Although this allowed direct comparisons of mood classification between Korean and American listeners, it inherently restricted Korean annota-

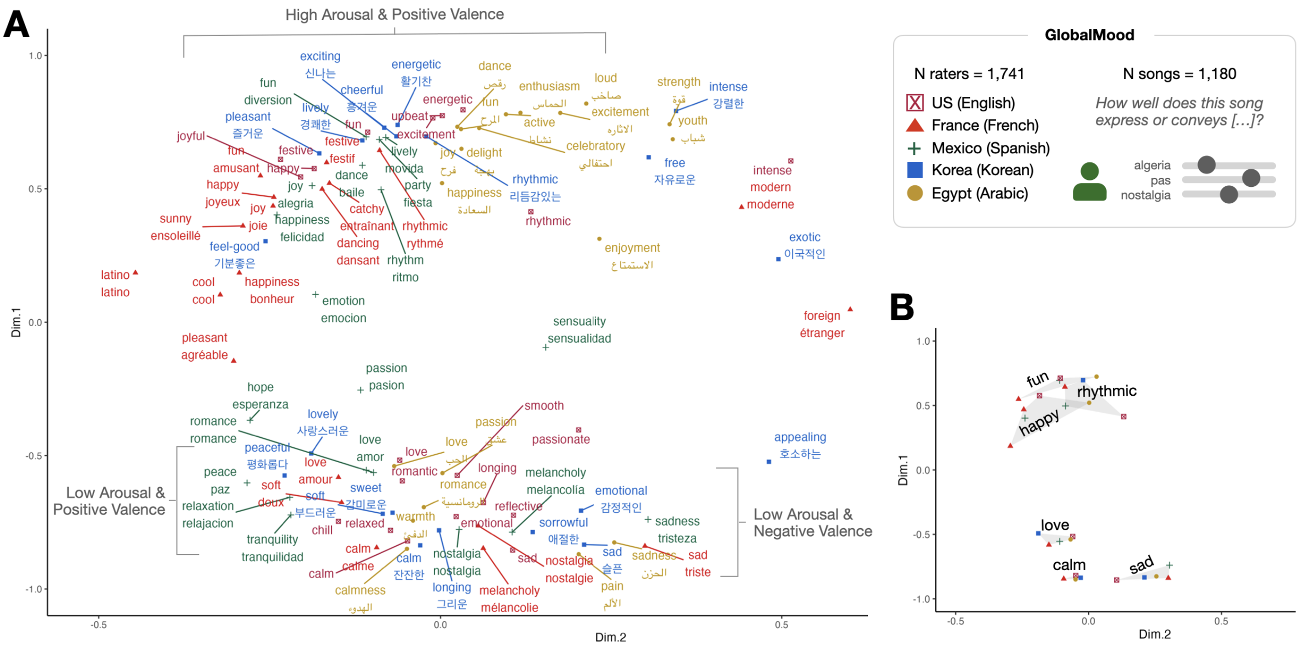

Figure 2. Association between mood terms across languages. **(A)** MDS visualization of mood terms based on mean ratings across the full song set. Terms positioned closer together exhibit similar rating patterns across songs, suggesting similar interpretations across languages. **(B)** Comparison of terms with direct translation equivalents across languages. The area size indicates the degree of semantic divergence in perception despite apparent translation equivalence.

tions to terms originally defiend within Western contexts.

More recently, Lee et al. [10] compiled a balanced set of American, Brazilian, and Korean songs and gathered mood annotations across nine categories, where annotators rated songs both from their own and the other two countries. They showed that certain mood terms like 'energetic' and 'sad' are highly consistent across cultures, while more abstract concepts like 'love' and 'dreamy' diverge considerably. Similar findings have been reported by other studies [11, 12], highlighting that when mood descriptors are imposed from one language onto another, important meanings can simply be 'lost in translation'.

In summary, while existing MER datasets and research have laid a solid groundwork, they remain limited by insufficient linguistic and cultural diversity. Because many are predominantly English-based and rely on top-down annotation strategies, they may overlook how people in other cultural contexts perceive mood of the music. As awareness of these limitations grows, there have been increasing calls for more inclusive approaches. Our work responds to this need by introducing a crowd-sourcing approach that organically elicits mood terms across languages, provided with a diverse, globally balanced set of musical examples.

2.2 Audio LLMs: the New Frontier in Music Tagging

Recent advances in multimodal large language models (LLMs) have opened promising avenues for downstream MIR tasks, including mood recognition. These models combine the reasoning capabilities of LLMs with audio perception systems (audio LLMs), enabling more flexible and nuanced music understanding than traditional classification approaches [21–23].

Models like MuLan [24] and MERT [21] have demonstrated potential for zero-shot music mood classification by embedding audio and natural language descriptions in a shared semantic space. However, comprehensive benchmarks such as the MuChoMusic [22] highlight a crucial limitation: these models rely heavily on language modality and do not attend sufficiently to audio, often failing with more nuanced audio examples for downstream MIR tasks. This limitation could be particularly critical for non-Western music and non-English mood descriptors, given that their training data are largely from Western contexts.

Similarly, closed-source models (e.g., Gemini) have shown promise in psychological textual analysis in multilingual contexts [25], while evaluations in specialized domains such as MIR remain scarce. The proprietary nature of their training data complicates thorough assessment of cross-lingual or cross-cultural performance. We aim to address these fundamental gaps by providing a large set of diverse, multilingual descriptors and annotations to support broader cross-cultural generalizability of audio LLMs.

3. METHOD

3.1 Participants

We recruited two independent sets of participants across the two stages of our data collection: **Stage 1** for emotion term elicitation (N = 778; see Section 4.1) and **Stage 2** for subsequent ratings on top 20 terms (N = 1,741; see Section 4.2). Participants had to be at least 18 years old, reside in the target country, and speak the target language as their primary language. Participants from the US were recruited through Prolific, while participants from the other four countries (France, Mexico, S. Korea, and Egypt) were recruited through the CINT platform. All participants provided informed consent under an approved protocol (see Section 6). Participants were instructed to wear headphones and take part in the study in a quiet environment. They had to pass a headphone screening task [26] and a language proficiency test [27] before being eligible for the main experimental task. Experiments were conducted in each participant's native language (English, French, Spanish, Korean, and Egyptian Arabic), instructions translated using GPT-4o. The repository accompanying this paper will include full task instructions and code to replicate all human experiments using the *PsyNet* framework [28].

3.2 Globally Representative Song Selection

To create a globally representative music dataset, we first used weekly YouTube top 100 music charts (year 2017-2023) from 59 countries, spanning six continents, ensuring representation from both Western and non-Western music. To ensure each country's charts reflected its distinct popular music, we excluded any track appearing in more than one country's chart. This left us with a *country-exclusive* pool of songs. From this pool, we sampled 20 songs per country, yielding 1,180 songs in total. Note that this diverse set is designed to capture a wide range of musical traditions and serve as a robust testbed for cross-cultural mood recognition. Each 15-second audio excerpt was trimmed from a random starting point in the full track, and then normalized at -5dB loudness.

In Stage 1 (Section 4.1), of the 1,180 full songs, we selected a subset of 180. For each country, we then added an additional 20 local songs (drawn from that country's pool) to construct a country-specific subset of 200 songs (for participants from that country). This step ensured that local participants encounter enough music strongly tied to their background, allowing them to elicit culturally specific mood descriptors. In Stage 2 (Section 4.2), the entire set of 1,180 songs was rated by participants to capture a comprehensive cross-cultural perspective.

3.3 Model Evaluation

Leveraging our novel GlobalMood dataset, we evaluated several recent multimodal and multilingual models capable of music understanding. Specifically, we assessed Google's Gemini models (*1.5 Flash*, *2.0 Flash*, and the latest *2.5 Pro*), a family of multimodal large language models capable of processing and reasoning across text and audio (but also image and video). We compared zero-shot and few-shot approaches, where the latter included 10 human-rated mood terms as examples.

Given that Gemini is closed-source, we also included CLAP (Contrastive Language-Audio Pretraining) [29] as an alternative, open-source model that learns joint audio-text embeddings. CLAP has demonstrated promise in MIR applications [30] and serve as the foundation for music-specific models like CLaMP [31]. Here, we conducted zero-shot evaluations through: (1) extracting audio embeddings from CLAP, (2) computing cosine similarities with text embeddings of mood terms, and (3) comparing these scores to human ratings.

We also fine-tuned CLAP on GlobalMood (train–test split = 1,000:180) to assess potential performance improvements. To preserve the continuous nature of our ratings, we represented each mood term in proportion to its mean rating (e.g., the term'calm', with a mean rating of 3.0, appeared three times in the text). This method retained the nuanced information in our soft labels rather than reducing them to binary categories. To improve generalizability, we created 10 augmented variations of each song through pitch shifting (range of ± 3 semitones), loudness adjustment (range of ± 15dB), and the addition of noise (amplitude of 0.005). Each augmented variant randomly included one or two of these modifications. The fine-tuned model checkpoints will also be publicly released.

Note that preliminary tests with recent multimodal models showed performance issues—Flamingo 2 [32] struggled with rating consistency and GPT-4o [33] failed to generate musical descriptions or ratings from audio alone—thus we excluded them from further analysis.

4. RESULTS

4.1 Bottom-up Term Elicitation Across Languages

4.1.1 Tagging pipeline

Many existing studies on music emotions rely on predefined taxonomies or web-scraped data that offer limited linguistic diversity [8, 17]. To overcome this limitation, we employed a bottom-up, participant-driven tagging method [11, 13, 14]. Specifically, we asked participants in each country to complete independent 'chains' of iterative annotations on each of 200 representative sample songs (see Section 3.2 for details of the stimuli).

Figure 1A illustrates one such chain: (i) The first participant annotates the song using single-word mood or emotion tags in their native language; (ii) The second participant rates the relevance of these tags (1–5 scale), flag irrelevant tags (e.g., genre- or lyrics-related rather than mood), add new tags as necessary; (iii) The third participant sees all tags from earlier participants and repeats these steps; (iv) This iterative process continues through ten participants per chain, systematically refining and validating mood terms.

To mitigate potential bias from earlier participants' contributions (i.e., priming effect), we conducted two parallel sets of annotation chains for each country.

4.1.2 Top emerging mood terms

Following the removal of tags flagged by more than two participants in a chain, our STEP-Tag process yielded an extensive, culturally specific lexicon of mood terms across languages: 644 unique terms in English, 528 in French, 870 in Spanish, 629 in Korean, and 283 in Arabic. To identify the most salient terms in each language, we calculated a composite score for every term by multiplying its frequency of occurrences across chains by its mean relevance rating. Higher scores indicate terms that participants frequently mentioned and consistently rated as highly relevant.

We consolidated closely related morphological variants (e.g., 'happy' and 'happiness' in English; gendered forms such as 'joyeux' and 'joyeuse' in French) manually with native speakers. This process was straightforward and unambiguous, and indeed, repeating this process with LLM (GPT-4o) provided nearly identical results. Figure 1B presents the resulting 20 highest-ranking mood tags per language, displaying both the original word and English translations to facilitate cross-cultural comparisons.

4.2 Large-scale Diverse Human Ratings

4.2.1 Cross-cultural ratings across the entire set

Having identified the top 20 mood terms for each language, we next gathered exhaustive ratings for all 1,180 songs. We recruited 1,741 new participants (see Section 3.1) who listened to the 15-second excerpts and rated how effectively each except conveyed a given mood term (1–5 scale). For each stimulus, participants evaluated seven randomly selected mood terms from the relevant language set. This systematic approach ensured that, on average, each song in each language received 8.38 (SD = 2.40) unique participant ratings, resulting in an extensive collection of 988,925 ratings spanning 1,180 songs across five languages.

4.2.2 Is 'happy' in my language the same 'happy' in your language?

To investigate differences in how each culture interprets these terms, we constructed 100 rating vectors (5 languages $\times$ 20 mood terms per language). Each vector was 1,180-dimensional, capturing the mean rating per term for each of the 1,180 songs. We then performed nonmetric multidimensional scaling (MDS) using correlation as the distance metric, projecting these vectors in a two-dimensional space. In this mood 'space,' terms that position close to one another—even those from different languages—reflect similar rating patterns across the musical examples, suggesting comparable mood or emotional interpretations across cultures.

Figure 2A visualizes this mood space. The terms cluster into two main regions: high arousal and high valence (e.g., *happy*, *energetic*, and *lively*; upper region of the figure) and low arousal that span positive valence (e.g., *peaceful*; bottom left) to negative (e.g., *sad*; bottom right). Notably, many translated "equivalents" appear close together, which might suggest a general cross-cultural consensus.

However, examining six commonly shared terms that have direct translations in at least four of the five languages (*fun*, *happy*, *rhythmic*, *love*, *sad*, and *calm*) revealed varying degrees of cross-cultural agreement (see Figure 2B). For each of these terms, we computed mean correlations across language pairs and additionally reported the within-country measurement error (r_{within}), calculated using split-half reliability with Spearman-Brown formula (i.e., inter-country agreement across terms).

The term *calm* showed the highest average agreement (r = 0.52 [0.49, 0.55]; r_{within} = 0.49 [0.38, 0.57]), followed by *fun* (r = 0.46 [0.39, 0.53]; r_{within} = 0.44 [0.27, 0.54]), *love* (r = 0.44 [0.41, 0.47]; r_{within} = 0.47 [0.33, 0.66]), *sad* (r = 0.41 [0.37, 0.45]; r_{within} = 0.43 [0.32, 0.58]), *rhythmic* (r = 0.38 [0.30, 0.45]; r_{within} = 0.43 [0.25, 0.53]), and notably *happy* (r = 0.37 [0.28, 0.45]; r_{within} = 0.45 [0.24, 0.59]).

Overall, considering within-country agreement (mean r_{within} = 0.43–0.48), most of these terms were comparable in their cross-country agreement (mean r = 0.39–0.52). However, despite being considered a basic universal human emotion [34], *happy*, exhibited a considerable gap between cross- and within-country agreement. This emphasizes the necessity of incorporating diverse cultural perspectives when modeling nuanced musical mood responses. Reliance on either dictionary translation or LLM-based translation alone could overlook important, context-specific nuances in mood perception—particularly relevant when building models for global audiences.

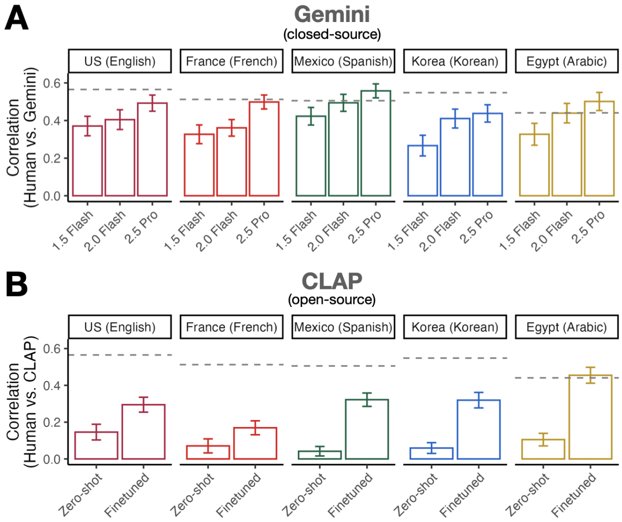

Figure 3. Correlations between human ratings and multimodal model predictions. **(A)** Gemini models with zero-shot prompting showing increase in performance with newer models. **(B)** CLAP models in zero-shot and fine-tuned scenarios showing how the use of multilingual annotations can substantially increase performance. Gray dashed lines represent split-half reliability of human ratings using the Spearman-Brown formula as baseline reference of correlations achieved between humans. Error bars indicate 95% CI of mean correlation across songs.

4.3 Human vs. Multimodal Models

Recent benchmarks have evaluated the capabilities of audio LLMs across various downstream MIR tasks, but these evaluations have also been restricted to English [22]. Our cross-cultural dataset provides a unique opportunity to assess their performance across multiple languages. We evaluated both closed-source (Gemini) and open-source (CLAP) models against our GlobalMood multilingual human ratings (see Section 3.3 for model details).

For Gemini (Figure 3A), we replicated our human study protocol by prompting the model with 'Rate from a scale of

1 to 5 how well this song expresses or conveys the emotion [...]' in each of the five native languages. We then evaluated how well the model's outputs predict human judgments. We observed consistent improvement with each new model version. The earliest model, *1.5 Flash*, demonstrated modest alignment with human judgments, achieving an average cross-country correlation of r = 0.34 (95% CI = [0.27, 0.42]). This model struggled particularly with Korean language ratings (r = 0.27 [0.21, 0.33]). The subsequent *2.0 Flash* version bridged this gap in Korean (r = 0.42 [0.36, 0.47]), and achieved a substantially higher overall cross-country correlation of r = 0.42 [0.36, 0.48]. The latest *2.5 Pro* model demonstrated another leap with a cross-country average correlation of r = 0.50 [0.45, 0.55]. Few-shot approach with 10 human-rated examples using 2.5 Pro did not improve the results (r = 0.47 [0.42, 0.52]). This consistent upward trajectory across model iterations provides compelling evidence that these systems are progressively developing more sophisticated, human-like capabilities for understanding musical emotions across diverse linguistic and cultural contexts.

This level of correlations between Gemini and human ratings are on par with algorithms specifically designed for estimating mood, such as Spotify's high-level feature estimation [10, 35]. Moreover, given the subjective nature of mood perception in music (where even humans often disagree), the latest Gemini model already reaches human-level performance, matching the theoretical upper bound defined by inter-rater human agreement (gray dashed lines in Figure 3).

For CLAP, an open-source alternative (Figure 3B), we evaluated both a zero-shot approach and a fine-tuned version trained on our GlobalMood dataset (see Section 3.3 for fine-tuning details). The zero-shot approach measured the similarity between the mood term's text embedding and the song's audio embedding. This zero-shot CLAP performed poorly (mean r = 0.08 [0.03, 0.13]), while fine-tuning with GlobalMood substantially improved the performance (mean r = 0.31 [0.19, 0.44]). As a control experiment, we also fine-tuned CLAP using a dataset where all non-English mood terms were first translated into English using an LLM without any musical context (instead of original mood terms collected from native speakers). This translation-based approach failed to improve performance (mean r = 0.13 [-0.05, 0.32]), showing that the performance gains from fine-tuning come from culturally-specific information rather than from mere increase in data volume.

Importantly, improvements through finetuning were most pronounced for non-English languages. Based on Fisher's z test for correlation comparisons, Arabic showed the largest increase from r = 0.11 to 0.46 (z = 0.39), followed by Spanish (from 0.04 to 0.32; z = 0.29) and Korean (from 0.06 to 0.32; z = 0.27). The least substantial increase was observed for French (from 0.07 to 0.17; z = 0.10). These observations demonstrate the promising potential of how native-language terms and ratings from our GlobalMood dataset can be used to generalize to new cultures and languages, when modeling MER systems.

5. DISCUSSION

We introduce GlobalMood, a novel cross-cultural benchmark dataset comprising mood descriptors and ratings collected from a large and globally diverse participant pool through a data-driven approach. Consistent with previous research in music perception [36–40] and music emotion [10–12, 41, 42], our findings highlight both cross-cultural similarities and differences. Specifically, we demonstrate that music mood descriptors across languages are broadly organized around clusters relating to valence and arousal [7, 34].

Relying on dictionary translations for music mood terms may face limitation (e.g. [41]), where cross-lingual agreement varied across terms. For instance, despite *happy* being basic emotion, it exhibited relatively large gap between cross- and within country agreement. This highlights the need for incorporating multilingual mood descriptors from diverse annotators. Expanding the dataset to include additional languages [27, 43] and a broader range of musical styles, including works from different historical periods, will be critical for developing more robust tools that can generalize across diverse cultural contexts. Our selection of popular songs often includes lyrics and participants' understanding of the lyrics may also have influenced their mood ratings. Future work could include instrumental music to better isolate mood judgments based solely on acoustic features.

Notably, fine-tuning CLAP on cross-cultural data significantly boosted performance in non-English contexts, highlighting the value of GlobalMood. These findings point to the potential for culturally-sensitive MER systems, moving beyond one-size-fits-all models. Future work could enhance this further through prompting, more diverse data augmentation techniques, and exploring other LLM architectures.

The tagging pipeline we adopt is quick to deploy across cultures and uses a bottom-up approach that helps minimize researcher bias. This offers an advantage over predefined mood lists, which often fail to capture the culturally grounded meanings of mood terms [44]. However, it has drawbacks: early participants can bias the tag pool, redundant or ambiguous tags require manual cleanup, and there's no theoretical guarantee that all tags reflect clear mood concepts (e.g., étranger in French). Unlike structured systems for crowdsourcing tags (e.g., TagATune [45]), our approach trades control for flexibility and cultural responsiveness.

Beyond MIR, the GlobalMood dataset and associated pipeline also offer timely implications for broader multimodal and cross-lingual research, particularly in Natural Language Processing (NLP) communities [46]. As NLP increasingly tackles multimodal tasks involving audio–text modeling and cross-cultural applications, our dataset may provide a useful benchmark for evaluating language–audio models across diverse contexts. The iterative annotation pipeline can also serve as an effective framework for collecting representative samples and annotations in other domains [13, 14, 28, 47].

6. ETHICS STATEMENT

We conducted our human experiment according to ethical best practices. All participants recruited via Prolific or CINT provided informed consent based on an approved protocol (Max Planck Ethics Council #202142). Participant data was collected anonymously (except for Prolific or CINT IDs), and all published data are fully anonymized. Models were accessed through commercial APIs, with fine-tuning performed locally. Code and training checkpoints will be provided in the published version. Conducting cross-cultural research requires sensitivity to diverse ethical considerations [48], and this study was designed accordingly. We acknowledge that machine learning models and training datasets may contain biases originating from participants, data selection processes, and the models themselves. Despite these potential biases, our study aims to address critical limitations and societal risks within current MIR technologies, which predominantly focus on English-language music and raters despite serving diverse global populations [44, 49, 50].

7. REFERENCES

[1] M. Zentner, D. Grandjean, and K. R. Scherer, "Emotions evoked by the sound of music: Characterization, classification, and measurement," *Emotion*, vol. 8, no. 4, pp. 494–521, 2008.

[2] A. S. Cowen, X. Fang, D. Sauter, and D. Keltner, "What music makes us feel: At least 13 dimensions organize subjective experiences associated with music across different cultures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, no. 4, pp. 1924–1934, Jan 2020.

[3] T. Eerola and P. Saari, "What emotions does music express? structure of affect terms in music using iterative crowdsourcing paradigm," *PLOS ONE*, vol. 20, no. 1, p. e0313502, Jan 2025.

[4] H. Tran, T. Le, A. Do, T. Vu, S. Bogaerts, and B. Howard, "Emotion-aware music recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 13, 2023, p. 13.

[5] Y. Sun, M. Kuo, X. Wang, W. Li, and Q. Bai, "Emotion-conditioned musiclm: Enhancing emotional resonance in music generation," in *2024 IEEE Congress on Evolutionary Computation (CEC)*, Jun 2024, pp. 1–8.

[6] P. N. Juslin and D. Västfjäll, "Emotional responses to music: The need to consider underlying mechanisms," *Behavioral and brain sciences*, vol. 31, no. 5, pp. 559–575, 2008.

[7] T. Eerola and J. K. Vuoskoski, "A review of music and emotion studies: Approaches, emotion models, and stimuli," *Music Perception: An Interdisciplinary Journal*, vol. 30, no. 3, pp. 307–340, 2012.

[8] X. Hu and J. S. Downie, "Exploring mood metadata: Relationships with genre, artist and usage metadata," in *International Society for Music Information Retrieval Conference*, Dec 2007.

[9] J. C. Jackson, J. Watts, T. R. Henry, J.-M. List, R. Forkel, P. J. Mucha, S. J. Greenhill, R. D. Gray, and K. A. Lindquist, "Emotion semantics show both cultural variation and universal structure," *Science*, vol. 366, no. 6472, pp. 1517–1522, 2019.

[10] H. Lee, F. Höger, M. Schönwiesner, M. Park, and N. Jacoby, "Cross-cultural mood perception in pop songs and its alignment with mood detection algorithms," in *The 22nd International Society for Music Information Retrieval Conference*, Nov 2021, pp. 366–373.

[11] E. Celen, P. van Rijn, H. Lee, and N. Jacoby, "Are expressions for music emotions the same across cultures?" Feb 2025, arXiv: arXiv:2502.08744.

[12] J. S. Gómez-Cañón, E. Cano, P. Herrera, and E. Gómez, "Joyful for you and tender for us: the influence of individual characteristics and language on emotion labeling and classification," in *Proceedings of the 21st International Society of Music Information Retrieval Conference (ISMIR)*. Montréal, Canada: Zenodo, Oct 2020.

[13] R. Marjieh *et al.*, "Words are all you need? language as an approximation for human similarity judgments," in *The Eleventh International Conference on Learning Representations*, Feb 2023.

[14] P. van Rijn, S. Mertes, K. Janowski, K. Weitz, N. Jacoby, and E. André, "Giving robots a voice: Human-in-the-loop voice creation and open-ended labeling," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–34.

[15] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Towards musical query-by-semantic-description using the CAL500 data set," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '07. New York, NY, USA: Association for Computing Machinery, Jul 2007, pp. 439–446.

[16] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PLOS ONE*, vol. 12, no. 3, p. e0173392, Mar 2017.

[17] J. Affolter and M. A. Rohrmeier, "Utilizing listener-provided tags for music emotion recognition: A data-driven approach," in *Proceedings of the 25th International Society for Music Information Retrieval Conference*. San Francisco, California, USA and Online: ISMIR, Nov 2024, pp. 547–554.

[18] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The MTG-Jamendo dataset for automatic music tagging," in *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*, Long Beach, CA, United States, 2019.

[19] A. Aljanaki, F. Wiering, and R. C. Veltkamp, "Studying emotion induced by music through a crowdsourcing game," *Information Processing & Management*, vol. 52, no. 1, pp. 115–128, Jan 2016.

[20] X. Hu, J. H. Lee, K. Choi, and J. S. Downie, "A cross-cultural study on the mood of K-POP songs," in *Proceedings of the 15th International Society for Music Information Retrieval Conference*. Taipei, Taiwan: ISMIR, Oct 2014, pp. 385–390.

[21] Y. Li *et al.*, "Mert: Acoustic music understanding model with large-scale self-supervised training," *arXiv preprint arXiv:2306.00107*, Apr 2024.

[22] B. Weck, I. Manco, E. Benetos, E. Quinton, G. Fazekas, and D. Bogdanov, "Muchomusic: Evaluating music understanding in multimodal audio-language models," in *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR 2024)*. San Francisco, California, USA and Online: ISMIR, Nov 2024.

[23] S. Doh, K. Choi, J. Lee, and J. Nam, "LP-MusicCaps: LLM-based pseudo music captioning," in *Proceedings of the 24th International Society for Music Information Retrieval Conference*, 2023.

[24] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. W. Ellis, "MuLan: A joint embedding of music audio and natural language," *arXiv preprint arXiv:2208.12415*, 2022.

[25] S. Rathje, D.-M. Mirea, I. Sucholutsky, R. Marjieh, C. E. Robertson, and J. J. Van Bavel, "GPT is an effective tool for multilingual psychological text analysis," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 121, no. 34, p. e2308950121, Aug. 2024.

[26] A. E. Milne, R. Bianco, K. C. Poole, S. Zhao, A. J. Oxenham, A. J. Billig, and M. Chait, "An online headphone screening test based on dichotic pitch," *Behavior Research Methods*, vol. 53, no. 4, pp. 1551–1562, 2021.

[27] P. van Rijn *et al.*, "Around the world in 60 words: A generative vocabulary test for online research," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2023.

[28] P. Harrison *et al.*, "Gibbs sampling with people," in *Advances in Neural Information Processing Systems*, 2020, pp. 10 659–10 671.

[29] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.

[30] J. Barnett, H. F. Garcia, and B. Pardo, "Exploring musical roots: Applying audio embeddings to empower influence attribution for a generative music model," *arXiv preprint arXiv:2401.14542*, Jan 2024.

[31] S. Wu, Z. Guo, R. Yuan, J. Jiang, S. Doh, G. Xia, J. Nam, X. Li, F. Yu, and M. Sun, "Clamp 3: Universal music information retrieval across unaligned modalities and unseen languages," 2025.

[32] S. Ghosh *et al.*, "Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities," 2025.

[33] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[34] P. Ekman, "Are there basic emotions?" *Psychological Review*, vol. 99, no. 3, pp. 550–553, 1992.

[35] D. Duman, P. Neto, A. Mavrolampados, P. Toiviainen, and G. Luck, "Music we move to: Spotify audio features and reasons for listening," *PLoS One*, vol. 17, no. 9, p. e0275228, Sep. 2022.

[36] N. Jacoby, E. A. Undurraga, M. J. McPherson, J. Valdés, T. Ossandón, and J. H. McDermott, "Universal and non-universal features of musical pitch perception revealed by sung reproduction," *Current Biology*, vol. 29, no. 19, pp. 3229–3243, 2019.

[37] S. A. Mehr, M. Singh, D. Knox, D. M. Ketter, D. Pickens-Jones, S. Atwood, C. Lucas, A. A. Egner, N. Jacoby, E. J. Hopkins *et al.*, "Universality and diversity in human song," *Science*, vol. 366, no. 6468, p. eaax0868, 2019.

[38] N. Jacoby, R. Polak, J. A. Grahn, D. J. Cameron, K. M. Lee, R. Godoy, E. A. Undurraga, T. Huanca, T. Thalwitzer, N. Doumbia *et al.*, "Commonality and variation in mental representations of music revealed by a cross-cultural comparison of rhythm priors in 15 countries," *Nature Human Behaviour*, vol. 8, no. 5, pp. 846–877, 2024.

[39] P. E. Savage, S. Brown, E. Sakai, and T. E. Currie, "Statistical universals reveal the structures and functions of human music," *Proceedings of the National Academy of Sciences*, vol. 112, no. 29, pp. 8987–8992, 2015.

[40] J. H. McDermott, A. F. Schultz, E. A. Undurraga, and R. A. Godoy, "Indifference to dissonance in native amazonians reveals cultural variation in music perception," *Nature*, vol. 535, no. 7613, pp. 547–550, 2016.

[41] A. S. Cowen, X. Fang, D. Sauter, and D. Keltner, "What music makes us feel: At least 13 dimensions organize subjective experiences associated with music across different cultures," *Proceedings of the National Academy of Sciences*, vol. 117, no. 4, pp. 1924–1934, 2020.

[42] T. Fritz, S. Jentschke, N. Gosselin, D. Sammler, I. Peretz, R. Turner, A. D. Friederici, and S. Koelsch, "Universal recognition of three basic emotions in music," *Current Biology*, vol. 19, no. 7, pp. 573–576, 2009.

[43] J. P. Niedermann, I. Sucholutsky, R. Marjieh, E. Celen, T. Griffiths, N. Jacoby, and P. van Rijn, "Studying the effect of globalization on color perception using multilingual online recruitment and large language models," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 46, 2023.

[44] D. E. Blasi, J. Henrich, E. Adamou, D. Kemmerer, and A. Majid, "Over-reliance on English hinders cognitive science," *Trends in Cognitive Sciences*, vol. 26, no. 12, pp. 1153–1170, 2022.

[45] E. L. M. Law, L. von Ahn, R. B. Dannenberg, and M. Crawford, "Tagatune: A game for music and sound annotation," in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, 2007, pp. 361–364.

[46] D. Hershcovich, S. Frank, H. Lent, M. de Lhoneux, M. Abdou, S. Brandl, E. Bugliarello, L. Cabello Piqueras, I. Chalkidis, R. Cui, C. Fierro, K. Margatina, P. Rust, and A. Søgaard, "Challenges and strategies in cross-cultural NLP," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6997–7013.

[47] D.-M. Huang, P. Van Rijn, I. Sucholutsky, R. Marjieh, and N. Jacoby, "Characterizing similarities and divergences in conversational tones in humans and llms by sampling with people," *arXiv preprint arXiv:2406.04278*, 2024.

[48] N. Jacoby, E. H. Margulis, M. Clayton, E. Hannon, H. Honing, J. Iversen, T. R. Klein, S. A. Mehr, L. Pearson, I. Peretz *et al.*, "Cross-cultural work in music cognition: Challenges, insights, and recommendations," *Music Perception*, vol. 37, no. 3, pp. 185–195, 2020.

[49] J. Henrich, S. J. Heine, and A. Norenzayan, "Most people are not weird," *Nature*, vol. 466, no. 7302, pp. 29–29, 2010.

[50] B. Amiri, N. Shahverdi, A. Haddadi, and Y. Ghahremani, "Beyond the trends: Evolution and future directions in music recommender systems research," *IEEE Access*, 2024.