CalmSet: A Domain-Specific Test Collection for Affective Music Retrieval for Children with ASD

Abhishek Karwankar
karwabhi@udel.edu
Computer and Information Sciences
University of Delaware
Newark, Delaware, USA

Liam Stapley
lstapley@udel.edu
Computer and Information Sciences
University of Delaware
Newark, Delaware, USA

Daniel Stevens
stevens@udel.edu
School of Music
University of Delaware
Newark, Delaware, USA

Matthew Louis Mauriello
mlm@udel.edu
Computer and Information Sciences
University of Delaware
Newark, Delaware, USA

Abstract

Information Retrieval (IR) increasingly relies on subjective, graded, and natural-language notions of relevance, motivating the development of reproducible test collections for ranking and recommendation. In affect-sensitive music domains, however, such resources with human-validated relevance signals remain scarce. We introduce CalmSet, a test collection for emotion-tagged music retrieval and recommendation in a therapeutic context for children with Autism Spectrum Disorder (ASD). CalmSet contains 432 modular music tracks instantiated from four purposefully composed base songs with controlled provenance, each formed by distinct combinations of seven active musical layers. Each track is annotated with ranked top-3 therapeutic intent labels and natural-language descriptions. Annotations are produced via a hybrid human-in-the-loop pipeline: CLAP proposes candidate intent labels, a large language model generates auxiliary semantic descriptions, and crowd workers provide ranked judgments without exposure to model outputs; final labels are aggregated using a Borda-based procedure. As initial baselines, we evaluate five one-vs-rest multi-label classifiers over CLAP audio embeddings, observing moderate micro-F1 scores (up to 0.60) but low exact-match accuracy (<0.10), while top-3 label overlap is substantially higher (Jaccard@3 up to 0.48), motivating graded-relevance evaluation. CalmSet supports both sparse (e.g., BM25) and dense audio–text retrieval models using therapeutic labels or natural-language descriptions as queries.

CCS Concepts

• **Information systems** → *Music retrieval.*

Keywords

Music Information Retrieval; Autism Spectrum Disorder; Dataset; Machine Learning; Emotion Annotation

SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808586

ACM Reference Format:
Abhishek Karwankar, Liam Stapley, Daniel Stevens, and Matthew Louis Mauriello. 2026. CalmSet: A Domain-Specific Test Collection for Affective Music Retrieval for Children with ASD. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3805712.3808586

1 Introduction

Information retrieval (IR) research increasingly addresses retrieval problems in which relevance is subjective, graded, and expressed in natural language, including recommendation, multimodal search, and domain-specific retrieval tasks [41]. This trend has motivated neural representation learning and multi-stage ranking pipelines that combine strong lexical baselines (e.g., BM25 [33]) with dense retrieval and neural reranking [17, 19, 30]. However, progress in such settings depends critically on well-specified test collections with reproducible evaluation protocols [29, 42].

Music Information Retrieval (MIR) has made substantial progress in modeling affect, similarity, and semantics in music [24], but most publicly available music emotion resources [16] were created for general-purpose listening contexts. As a result, they often conflate emotion with genre/style cues, and their annotations (e.g., tags or broad listener impressions) can be difficult to interpret as retrieval relevance in affect-sensitive applications. Therapeutic music for children with Autism Spectrum Disorder (ASD) exemplifies a high-stakes, affect-sensitive retrieval setting [36]. These children may exhibit atypical auditory processing and heightened sound sensitivities where specific sounds can trigger distress or sensory overload rather than regulation [4, 26]. In this context, music is intentionally selected to support regulation and engagement [28, 52]; small variations in rhythm, texture, or instrumentation can substantially change perceived affect [10, 28]. These constraints expose a domain mismatch: retrieval and recommendation models trained on generic corpora may not generalize when relevance depends on subtle affective intent rather than coarse mood labels.

Existing datasets for Music Emotion Recognition (MER), including PMEmo [56], DEAM [3], XMusic [44], and GlobalMood [22], have driven progress in affect modeling [54]. Recently, contrastive audio–language models such as CLAP [7] and AST [9] enabled zero-shot annotation and cross-modal retrieval by aligning audio and text

representations [53]. Yet in specialized domains, zero-shot predictions and generic emotion vocabularies can be noisy or ambiguous, limiting their use as reliable relevance signals for IR evaluation.

To address this gap, we introduce **CalmSet**, a crowdsourced test collection for emotion-conditioned music retrieval and recommendation in a therapeutic domain. CalmSet is derived from a controlled, modular music corpus purposefully composed for children with ASD [18, 27]. Each song is constructed from seven rhythmic and instrumental layers, each with up to three alternatives that can be combined independently, yielding a structured musical space with over 10K possible variations. From this corpus, we curate a subset of 432 tracks (~2.5 minutes each) that spans diverse affective intent while maintaining consistent production quality. Critically, while the 432 tracks derive from four base songs, each was purposefully designed with distinct timbral, stylistic, and motivic profiles across seven audio layers and fifteen sublayers, yielding genuine acoustic and affective diversity rather than superficial variation [18, 27]. Unlike scraped music collections, controlled provenance reduces issues arising from genre and recording variability, making CalmSet suitable for benchmarking retrieval models under affect-sensitive constraints.

Our annotation pipeline integrates machine learning, natural language generation, and crowd-sourced validation. We first apply CLAP to perform zero-shot audio classification, assigning cosine similarity scores across eight emotional tags defined in consultation with composers. The top three tags with the highest scores are used to represent each track's predicted emotions, and GPT-4o generates natural-language descriptions conditioned on these tags. Although CalmSet targets a specialised therapeutic context, the annotation task was deliberately scoped to emotional perception rather than clinical assessment, a distinction validated through iterative discussion with music composers on our research team and prior work available on this topic [40]. Since emotional responses to music are grounded in general auditory perception, crowd workers are well-positioned to provide meaningful affective judgements without requiring domain expertise in ASD; the specialised knowledge of our collaborators instead informed the tag vocabulary and corpus design upstream of the annotation stage. To validate these outputs, we recruited 16 qualified workers from a pool of 50 on Amazon Mechanical Turk, each of whom passed a gold-standard recruitment task. Every song was annotated by three workers, who were not shown CLAP predictions but were asked to rate their agreement with the generated descriptions and to rank their top three emotion tags. Final labels were aggregated using Borda count [8], with CLAP scores serving as tie-breakers when worker agreement was high but not unanimous. This design supports IR-style evaluation where relevance is graded and multiple interpretations may coexist [43]. CalmSet is publicly released with documentation, metadata schemas, and preprocessing scripts to support use by MIR and IR researchers, and was collected under approved institutional review board protocol (Protocol 20973251). The dataset can be accessed at: https://www.kaggle.com/datasets/sensifylab/calmset-music-for-children-with-asd/data. The code is available at: https://github.com/Sensify-Lab/CalmSet.

By providing a domain-specific benchmark with controlled provenance and crowd-sourced validations, CalmSet enables IR research on affect-aware music retrieval, recommendation, and cross-modal search in therapeutic and other affect-sensitive contexts. We contribute: (i) CalmSet, a dataset of 432 modular therapeutic music tracks with ranked emotion annotations and natural-language descriptions, designed to support reproducible evaluation of emotion-conditioned retrieval, and (ii) present a hybrid annotation pipeline that combines weak supervision with human ranking judgments and deterministic aggregation, yielding interpretable, rank-aware labels suitable for IR benchmarking.

2 Related Work

Here, we review (i) musical datasets used for MER, and (ii) modeling and annotation approaches that support emotion-aware retrieval and recommendation.

2.1 Music Emotion Recognition Datasets

MER has been a long-standing problem in MIR, with numerous datasets proposed to support benchmarking and model development [16, 54]. Early MER datasets primarily relied on categorical mood annotations. For example, the MIREX Mood Dataset [6] and CAL500 [45] provided song-level emotion labels for Western popular music using expert raters or undergraduate annotators. Subsequent extensions, such as CAL500exp [50], introduced fragment-level labels, enabling finer-grained temporal modeling. Other datasets, including the 4Q Emotion Dataset [31], mapped music into predefined emotion clusters or valence–arousal quadrants.Emotify [2] and EMOPIA [12] employ categorical labels derived from the Geneva Emotional Music Scales (GEMS) [25] or Russell's circumplex model [32]. In parallel, several datasets moved toward continuous dimensional representations, collecting time-varying arousal and valence ratings. The Yang-Dim dataset [55], MoodSwings [20], and MoodSwings-Turk [40] exemplify this approach, enabling dynamic emotion modeling over musical excerpts. Large-scale resources such as DEAM [3] and PMEmo [56] further expanded the field with thousands of tracks annotated via crowdsourcing, typically sampled at 1–2 Hz. Recent trends emphasize scale and diversity. XMusic [44] provides over 100,000 MIDI files, while GlobalMood [22] offers cross-cultural benchmarking. Weakly supervised datasets like the Million Song Dataset [5], MuSe [1], and Music4All [35] leverage listener-generated tags or platform-derived features (e.g., Spotify valence or danceability) to support large-scale modeling and retrieval. However, these resources are intended for entertainment contexts and reflect passive listener impressions, offering limited interpretability for sensitive applications.

CalmSet addresses this gap by introducing a dataset composed for therapeutic contexts involving children with ASD. Unlike prior datasets that capture perceived emotion during passive listening, CalmSet encodes composed emotional intent with interpretable annotations and agreement measures, enabling benchmarking for emotion-aware retrieval, recommendation, and representation learning in settings where reliability and interpretability are critical.

2.2 Emotion-Aware Music Retrieval

Traditional MER approaches map low-level audio features to categorical moods or arousal–valence dimensions using supervised learning [3, 54]. Deep learning has significantly expanded the MER

landscape. Transformer-based architectures, such as the Audio Spectrogram Transformer (AST) [9], and contrastive audio–language models, such as CLAP [7], enable scalable representation learning and zero-shot emotion annotation by aligning audio with natural-language concepts [53]. These models have become foundational tools for MER, cross-modal retrieval, and large-scale music tagging.

MER involves indexing tracks using dense representations, retrieving them via textual or intent-based queries, and evaluating them using standard metrics. Classic probabilistic ranking functions such as BM25 [33] and learning-to-rank methods [23] motivate strong baselines for text-to-music and emotion-conditioned retrieval. Recently, neural IR, including dense retrieval [17] and late-interaction rerankers [19, 30], suggest multi-stage pipelines in which cross-modal encoders retrieve candidates, and more expensive models refine rankings. Following benchmarks like MS MARCO [29] and BEIR [42], CalmSet serves as both an MER dataset and a IR benchmark for therapeutic, emotion-aware retrieval.

Beyond classification, MER increasingly supports emotion-aware retrieval and recommendation. Cross-modal systems match natural-language queries with audio embeddings to enable text-to-music search and emotion-aware recommendations [7, 11]. This enables non-experts to use everyday language for queries like 'soothing music for transitions' or 'stimulating tracks for activities'. In parallel, music recommender systems have begun to move beyond entertainment toward well being and accessibility, incorporating emotional cues alongside contextual signals such as time of day, activity, or user engagement [37]. At the same time, recent work highlights challenges associated with annotation noise and domain mismatch, especially when zero-shot or weakly supervised models are applied outside their training distributions. To address these issues, hybrid annotation strategies that combine automated predictions with targeted human validation have gained traction [14, 15, 51]. These hybrid pipelines balance scalability with reliability, improving label consistency and interpretability for benchmarking.

CalmSet integrates zero-shot CLAP predictions, descriptions, and human agreement scores in a hybrid annotation pipeline that balances computational scalability with validation. This yields annotations that are both machine-readable and interpretable, supporting downstream tasks such as MER benchmarking, emotion-conditioned retrieval, explainable recommendation, and studies of weak-to-strong supervision. By coupling modern audio–text models with domain-specific human validation, CalmSet complements existing MER datasets and enables research at the intersection of emotion, retrieval, and music therapy.

3 The CalmSet Dataset

Building on prior work introducing *uCue*, a tangible musical interface with a custom modular music library for therapeutic contexts [18, 27], we formalize this library as the *CalmSet* dataset. CalmSet comprises children's songs structured using a standardized template of seven interactive layers: *melody*, *harmony*, *countermelody*, *bass line*, *percussion*, *bass drum*, and *ambient sound*. Each layer can be independently activated, muted, or cycled through variants using a custom controller designed for children with ASD. Across the dataset, each song includes multiple variants per layer (e.g., three melodic variants, three countermelodies, two harmony layers, two bass lines, and three ambient textures), enabling a large combinatorial space of valid musical renditions. This modular structure supports both experimental control and expressive flexibility, allowing compositions to be adapted to different therapeutic goals such as calming, stimulation, or regulation. Within *uCue*, children explore and shape these layers interactively, constructing personalized musical experiences rather than consuming fixed recordings. All compositions were produced using a professional digital audio workflow, combining MIDI-based composition, live instrumental recording, and multitrack mixing in a DAW environment. A diverse palette of synthesizers, samplers, effects, and sound libraries was used to ensure that each song maintains a distinct sonic identity while remaining musically coherent.

Layer Design. Melodic layers vary systematically in timbre, register, and articulation, enabling controlled comparisons of listener preference across contrasting sound qualities. Countermelodies introduce increasing rhythmic activity and a wider registral span across three variants, allowing the study of tolerance for rhythmic and melodic complexity. Harmony layers provide two levels of stimulation: a sustained, diatonic texture designed for low-arousal contexts and a more rhythmically and harmonically active alternative for expressive exploration. Bass layers include a harmonically aligned foundation and a more rhythmically active variant, while percussion and bass drum layers provide optional rhythmic grounding with subtle variation. Ambient layers consist of contrasting environmental soundscapes, enabling investigation of how non-musical context interacts with musical structure.

Aesthetic Goals and Dataset Generation. All compositions were set at slower-than-typical tempos to accommodate neural processing differences, while preserving optional rhythmic drive through selected layers. To ensure both musical coherence and systematic variability, countermelodies and harmony layers were composed to recombine meaningfully with the primary melody. We enumerated all valid layer combinations using a Python-based generation script, yielding a large set of unique renditions suitable for both therapeutic deployment and computational analysis. As a result, CalmSet functions simultaneously as an engaging therapeutic music library and a controlled dataset for studying music interaction, preference, and affect-aware retrieval.

4 CalmSet Annotation Methodology

We design CalmSet as a crowd-sourced, human-validated test collection for emotion-conditioned music retrieval and recommendation. Our three-stage pipeline prioritizes reproducible relevance signals over single-label assignments. First, we use CLAP to generate a candidate set of emotion tags for each clip, which serves as a lightweight semantic prior and (later) a deterministic tie-breaker when human evidence is ambiguous. Second, we use an LLM to generate short natural-language descriptions conditioned on candidate tags; these descriptions serve as auxiliary metadata to support downstream retrieval baselines and analysis. Third, we collect human judgments on MTurk: workers provide a ranked top-3 list of emotion tags for each clip, and we aggregate these rankings with an agreement-weighted Borda procedure to obtain graded, rank-aware labels. Throughout, workers never observe the CLAP predictions;

Algorithm 1 CLAP-based Emotion Inference

Require: Audio clips $\mathcal{X} = \{x_k\}_{k=1}^{N}$, emotion phrases $\mathcal{P} = \{p_i\}_{i=1}^{8}$, CLAP audio encoder f_a, CLAP text encoder f_t
Ensure: CLAP top-3 emotion tags $\hat{\mathcal{E}}_k^{\text{CLAP}}$ and similarity scores $\hat{\mathcal{S}}_k^{\text{CLAP}}$

for $i \leftarrow 1$ **to** 8 **do**
  $\mathbf{b}_i \leftarrow f_t(p_i)$ {Text embeddings}
end for
for $k \leftarrow 1$ **to** N **do**
  $\mathbf{a}_k \leftarrow f_a(x_k)$ {Audio embedding}
  for $i \leftarrow 1$ **to** 8 **do**
    $s_{k,i} \leftarrow \dfrac{\mathbf{a}_k \cdot \mathbf{b}_i}{\|\mathbf{a}_k\| \, \|\mathbf{b}_i\|}$
  end for
  $\hat{\mathcal{E}}_k^{\text{CLAP}} \leftarrow \textsc{TopK}(s_{k,i}, K = 3)$
  $\hat{\mathcal{S}}_k^{\text{CLAP}} \leftarrow \textsc{TopKScores}(s_{k,i}, K = 3)$
end for
return $\hat{\mathcal{E}}_k^{\text{CLAP}}, \hat{\mathcal{S}}_k^{\text{CLAP}}$

Algorithm 2 LLM-based Song Description Generation

Require: CLAP top-3 tags $\hat{\mathcal{E}}_k^{\text{CLAP}}$, LLM $g(\cdot)$, fixed prompt template π
Ensure: GPT-generated description d_k

for $k \leftarrow 1$ **to** N **do**
  Construct prompt π_k using $\hat{\mathcal{E}}_k^{\text{CLAP}}$
  $d_k \leftarrow g(\pi_k)$ {2–3 sentence description}
end for
return $\{d_k\}_{k=1}^{N}$

CLAP is only used for tie-breaking during aggregation, to limit model anchoring while keeping the pipeline reproducible.

4.0.1 CLAP-based Emotion Inference. We first apply CLAP to perform zero-shot emotion inference on each therapeutic music clip. We encode 16 kHz audio clips and descriptive emotion phrases (see the code on GitHub for the descriptions) into a joint CLAP embedding space. Cosine similarity between audio and text embeddings is computed to quantify semantic alignment (See algorithm 1). For each clip, the top three emotion tags with the highest similarity scores are retained as the CLAP-predicted emotional profile. These predictions serve as an initial semantic prior and are later used only for tie-breaking during human annotation aggregation.

4.0.2 LLM-based Description Generation. To improve interpretability and support downstream therapeutic use, we convert the CLAP-predicted emotion tags into readable song descriptions using a large language model (code on Github for the prompt). For each song, the top three CLAP emotion tags are injected into a structured prompt designed from a music-therapy perspective (See algorithm 2). The model generates a concise textual summary describing the song's emotional qualities, musical characteristics, and potential therapeutic benefits. All parameters are fixed to ensure consistency.

Algorithm 3 Agreement-weighted Aggregation and Tie-breaking

Require: Worker rankings $r_{k,j}$, agreement scores $a_{k,j}$, CLAP tags $\hat{\mathcal{E}}_k^{\text{CLAP}}$, CLAP scores $\hat{\mathcal{S}}_k^{\text{CLAP}}$, $\alpha = 0.25$, $w_{\min} = 0.25$
Ensure: Final top-3 emotion tags $\hat{\mathcal{E}}_k^{\text{final}}$

for each song k **do**
  Initialize $S_k(t) \leftarrow 0$ for all tags t
  for each worker j **do**
    $w_{k,j} \leftarrow \max\{w_{\min}, 1 + \alpha a_{k,j}\}$
    Assign Borda points $(3, 2, 1)$ weighted by $w_{k,j}$
  end for
  Rank tags by $S_k(t)$
  Compute mean agreement $\bar{a}_k$
  if tie near top-3 **then**
    if $\bar{a}_k \geq 0$ **then**
      Resolve via CLAP order → CLAP score → worker frequency → alphabetical
    else
      Resolve via worker frequency → alphabetical
    end if
  end if
  Select top-3 tags as $\hat{\mathcal{E}}_k^{\text{final}}$
end for
return $\hat{\mathcal{E}}_k^{\text{final}}$

4.0.3 Crowdsourced Annotation and Final Label Aggregation. We finalize emotion annotations with crowdsourced relevance judgments collected via Amazon Mechanical Turk. Each track is annotated by three independent workers, who (i) select their top three emotion tags based on their own interpretation of the audio and (ii) rate the semantic coherence of the song description generated from algorithm 2. Workers remain blind to CLAP predictions, providing independent rankings and agreement scores to prevent model anchoring. Worker-provided rankings are aggregated using an agreement-weighted Borda count, in which votes are softly weighted by self-reported agreement with the description (Algorithm 3). Ties are resolved deterministically, incorporating CLAP only when human agreement is non-negative. This design emphasizes scalable, cost-effective relevance estimation while maintaining a clear separation between model inference and human annotation.

4.0.4 Results on human-annotated songs (n=432). Table 1 summarizes the aggregation outcomes. Roughly one-third of songs (33.1%) contained ties across workers' votes. In 82 cases (18.9%), these ties were resolved using CLAP, while in 61 cases (14.1%) they were resolved solely by worker frequency or fallback rules. On average, the agreement score across songs was positive ($\bar{a} = 0.82$, scale of [-2 to 2]), indicating that workers agreed with the CLAP-generated descriptions most of the time. This agreement allowed CLAP to participate in tie-breaking in over 91% of cases.

4.0.5 Comparison with CLAP. We use Jaccard similarity (J) to quantify alignment between human and CLAP tags:

$$J = \frac{|\,\text{Top3}^{\text{final}} \cap \text{Top3}^{\text{CLAP}}\,|}{|\,\text{Top3}^{\text{final}} \cup \text{Top3}^{\text{CLAP}}\,|}.$$

Table 1: Label aggregation for the human-annotated subset

Category	Count	Percentage
Total songs	432	100%
Songs with ties	143	33.1%
Resolved with CLAP	82	18.9%
Resolved without CLAP	61	14.1%

The average overlap was 0.35 (median 0.200), indicating that CLAP alone is insufficient to capture labels, but in over half of the tie cases it provided a useful prior that complemented human evidence. Our aggregation method balances three sources of evidence: (i) relative ranking from multiple workers, (ii) their confidence in the GPT description, and (iii) CLAP as a domain-informed tie-breaker. This yields a transparent, reproducible, and human-grounded set of annotations. Importantly, CLAP only intervenes when workers' agreement with the description is positive, ensuring that model priors do not override human disagreement.

4.0.6 Dataset Characterization. CalmSet comprises 432 therapeutic music tracks, each annotated with a ranked top-3 set of therapeutic intent labels. Label frequencies exhibit moderate imbalance: the most prevalent labels are *Sensory-Calming* (272 tracks), *Playful* (248), and *Soothing* (222), while *Anxiety-Reduction* (48) and *Transitional* (78) form a long tail. At rank `top1`, *Sensory-Calming* appears most frequently (131 tracks), followed by *Playful* (111) and *Soothing* (75), whereas lower-frequency labels are more evenly distributed across `top2` and `top3`. Label co-occurrence is common, reflecting overlapping therapeutic intents. Frequent label pairs include *Sensory-Calming–Soothing* (150 tracks), *Sensory-Calming–Playful* (119), and *Sensory-Calming–Grounding* (116). These patterns motivate the use of graded relevance rather than binary judgments.

5 Experiments: Retrieval Task and Evaluation Protocol

Labels are treated as queries and audio files as documents, using the top-3 human-validated labels for graded relevance (rel $\in \{3, 2, 1\}$).

Emotion-to-music retrieval task. We formulate emotion-to-music retrieval following standard IR conventions. Each of the eight therapeutic intent labels (*Anxiety-Reduction*, *Focusing*, *Grounding*, *Playful*, *Sensory-Calming*, *Soothing*, *Stimulating*, and *Transitional*) is treated as a query over the full collection of 432 tracks. Relevance is defined using ranked annotations: tracks annotated with the query label at `top1`, `top2`, or `top3` receive graded relevance scores of rel = 3, 2, 1, respectively, and rel = 0 otherwise. This yields 1,296 graded relevance assignments and supports gain-based evaluation using metrics such as nDCG and MAP [13, 34, 46]. The number of relevant documents per query ranges from 48 to 272, motivating macro-averaged evaluation across queries.

Evaluation protocol. All experiments follow a test-collection-style evaluation protocol common in information retrieval, in which retrieval methods are evaluated over the full collection using fixed queries and relevance judgments rather than train–test splits [47, 49]. Each of the eight therapeutic intent labels is treated as a query, and retrieval effectiveness is reported as macro-averages across queries to ensure equal weighting of intents. Statistical significance is assessed via paired tests, with 95% bootstrap confidence intervals computed over per-query scores [38, 39].

Table 2: Macro-averaged retrieval performance for emotion-to-music retrieval on CalmSet at $k = 50$.

Method	nDCG@50	MAP@50	Recall@50
Random	0.216	0.417	0.115
BM25 (GPT descriptions)	0.261	0.494	0.143
CLAP label-based	**0.293**	**0.540**	**0.162**

5.1 Baselines

We evaluate three representative methods: (1) Random, a lower-bound baseline averaged over five seeds; (2) BM25, a sparse lexical baseline using keyword queries over GPT-generated descriptions; and (3) CLAP, a dense retrieval baseline that ranks tracks based on the similarity between audio embeddings and therapeutic intent labels. All models use deterministic tie-breaking for reproducibility.

5.2 Metrics

We report standard ranking metrics used in graded retrieval. Ranking quality is primarily measured using nDCG@k, which accounts for graded relevance and rank position, emphasizing the retrieval of highly relevant tracks early in the ranked list. We additionally report MAP@k, which evaluates early precision by rewarding rankings that place relevant tracks near the top, and Recall@k, which measures coverage by quantifying how many relevant tracks are retrieved within the top-k results. MAP@k and Recall@k are computed using binarized relevance (relevant if rel > 0). All metrics are reported at $k = 50$ and macro-averaged across queries.

5.3 Results

Table 2 shows retrieval performance. CLAP achieved the highest effectiveness (nDCG@50 = 0.293), outperforming BM25 and the random baseline. MAP@50 increases from 0.417 (random) to 0.494 (BM25) and 0.540 (CLAP), while Recall@50 improves from 0.115 (random) to 0.143 (BM25) and 0.162 (CLAP). Bootstrap confidence intervals over queries yield a 95% CI of [0.175, 0.425] for CLAP nDCG@50 and [0.169, 0.367] for BM25, indicating overlapping consistently higher performance for CLAP. A paired test over queries comparing CLAP to random ranking yields a t-statistic of 2.01 ($df = 7$), indicating a consistent improvement over chance.

Given the limited number of queries, we complement aggregated metrics with a per-query breakdown in Table 3 to ensure no single query disproportionately drives overall results [48]. Performance varies substantially across intents. *Playful* and *Sensory-Calming* yield the highest CLAP nDCG@50 scores (0.576 and 0.525), suggesting their affective character is well-captured by audio embeddings. By contrast, *Anxiety-Reduction* and *Transitional* are consistently the weakest queries across all methods (CLAP nDCG@50 of 0.074 and 0.107 respectively), reflecting the greater semantic ambiguity and context-dependence of these intents. Notably, BM25 outperforms CLAP on nDCG@50 for *Stimulating* (0.247 vs. 0.224) and *Transitional* (0.150 vs. 0.107), and on *Anxiety-Reduction* (0.084 vs. 0.074),

Table 3: Per-query retrieval performance across the eight therapeutic intent labels at $k = 50$.

	Random			BM25			CLAP		
Query	nDCG	MAP	Rec	nDCG	MAP	Rec	nDCG	MAP	Rec
Anxiety-Reduction	0.082	0.170	0.092	**0.084**	**0.278**	0.146	0.074	0.246	**0.188**
Focusing	0.132	0.315	0.124	0.159	0.382	0.146	**0.161**	**0.412**	**0.175**
Grounding	0.215	0.481	0.120	0.191	**0.527**	**0.132**	**0.234**	0.420	0.111
Playful	0.357	0.644	0.121	0.465	0.598	0.129	**0.576**	**0.958**	**0.153**
Sensory-Calming	0.387	0.645	0.110	0.515	0.794	0.136	**0.525**	**0.802**	**0.147**
Soothing	0.286	0.538	0.112	0.280	0.633	0.153	**0.447**	**0.852**	**0.189**
Stimulating	0.163	0.326	0.118	**0.247**	**0.513**	0.147	0.224	0.431	**0.169**
Transitional	0.102	0.218	0.123	**0.150**	**0.227**	0.154	0.107	0.200	**0.167**

indicating that for these intents lexical cues in GPT-generated descriptions are more discriminative than audio embeddings alone. Conversely, CLAP holds a large advantage on *Soothing* (0.447 vs. 0.280) and *Playful* (0.576 vs. 0.465), where timbral and rhythmic features are more informative than text. Recall@50 remains low across all queries (maximum 0.189 for *Soothing* under CLAP), confirming that relevant tracks are broadly distributed in rankings regardless of method, and that the benchmark is far from saturated.

Retrieval performance differs across emotion labels for both text-based and audio-based methods. Intents with clear and easily recognizable characteristics (e.g. Sensory-Calming and Playful), are retrieved more reliably, whereas more subtle or situational intents (e.g. Anxiety-Reduction and Transitional), are harder for current models to capture. Across all labels, recall remains low, meaning relevant tracks are spread throughout the ranked results rather than appearing near the top. This suggests that existing text and audio representations struggle to fully capture the range of music that may be appropriate for a given therapeutic intent. These findings indicate that CalmSet contains meaningful affective patterns across both textual and audio information, while still posing a substantial challenge for current retrieval models, making it well-suited as a benchmark for affect-aware music retrieval.

5.4 Benchmarking MER Models

5.4.1 Setup. To establish baselines for our dataset, we evaluated five multi-label classifiers trained on CLAP audio embeddings: *Logistic Regression*, *Linear SVM*, *Random Forest*, *Multi-Layer Perceptron*, and *KNN*. We use a One-vs-Rest (OvR) strategy, which decomposes the task into binary classifiers. We split the data 70%/10%/20% for train/validation/test, stratified by primary label. We compare a default 0.5 threshold against validation-tuned per-class F1 thresholds.

5.4.2 Evaluation metrics. We report micro- and macro-averaged F1 and Jaccard similarity, and subset accuracy (exact match). We additionally compute **Top-3 overlap**, measuring Jaccard, precision, and recall between predicted and reference top-3 labels.

5.4.3 Main results. Table 4 summarizes performance on the test split. Across models, micro-F1 ranges from 0.57–0.60 and macro-F1 from 0.41–0.47. Logistic Regression achieved the strongest default test performance (micro-F1 = 0.602, macro-F1 = 0.471, micro-Jaccard = 0.431). Threshold calibration improved some models (e.g., KNN micro-F1 = 0.617, Random Forest = 0.603), though gains were inconsistent. Low subset accuracy (<0.10) reflects the difficulty of exact multi-label prediction in subjective affect domains.

Table 4: Test set results across baseline models. Default = 0.5 threshold. Calibrated = thresholds tuned on validation.

	Default (0.5)			Calibrated		
Model (OvR)	Micro-F1	Macro-F1	Micro-J	Micro-F1	Macro-F1	Micro-J
LogReg	**0.602**	**0.471**	**0.431**	0.574	0.482	0.403
Linear SVM	0.572	0.461	0.401	0.571	0.499	0.400
Random Forest	0.600	0.436	0.429	0.603	0.492	0.431
MLP	0.579	0.408	0.407	0.589	0.475	0.417
KNN	0.602	0.407	0.431	**0.617**	**0.507**	**0.446**

5.4.4 Top-3 overlap. Given top-3 annotations per song, Top-3 overlap provides an intuitive measure of alignment with human perception. KNN leads (mean Jaccard@3 = 0.484, P@3 = 0.613, R@3 = 0.613), closely followed by Logistic Regression (Jaccard@3 = 0.460). Overall, models retrieve at least one top-3 tag in over half of cases, demonstrating the utility of CLAP embeddings for therapeutic MER. These results highlight promise and limitations. CLAP-based classifiers show moderate alignment with human labels, particularly in Top-3 overlap, yet exact-match accuracy remains low due to multi-label complexity and subjective affect perception.

6 Discussion, Limitations, and Future Work

CalmSet is a domain-specific benchmark for affect-aware music retrieval and emotion recognition, emphasizing representation, ranking, and graded subjective relevance rather than therapeutic efficacy. Across baselines, CLAP-based label retrieval consistently outperforms BM25 over GPT-generated descriptions and random ranking, indicating that general-purpose audio–text representations capture coarse therapeutic intent more effectively than lexical cues alone. However, absolute performance remains modest, particularly in recall, suggesting relevant tracks are dispersed rather than concentrated at the top of rankings. This effect is most pronounced for semantically subtle or context-dependent intents (e.g., *Anxiety-Reduction* and *Transitional*), which lack distinctive acoustic or lexical signatures compared to more concrete intents such as *Sensory-Calming* or *Playful*. A natural question is whether these modest scores reflect the difficulty of affect-aware retrieval in this domain, or instead expose limitations specific to CLAP's representational capacity. While our baselines do not include a CLAP-free audio feature set, such as hand-crafted MFCCs or alternative neural embeddings, two observations support the former interpretation. First, BM25, which operates entirely over text and shares no features with CLAP, exhibits a similar pattern of difficulty: strong on *Playful* and *Sensory-Calming*, weak on *Anxiety-Reduction* and *Transitional*. This convergence across modalities suggests the difficulty is task-intrinsic rather than CLAP specific. Second, the intents that challenge both methods are those with the greatest semantic ambiguity and context-dependence, properties that would challenge any single-representation system regardless of backbone. This remains an open empirical question, and encourages future work to evaluate alternative audio representations on CalmSet to disentangle model limitations from benchmark difficulty. These findings suggest that emotional-intent access is better framed as ranking rather

than single-label prediction. In CalmSet, tracks are associated with multiple intents to varying degrees, making evaluation based on ranked relevance more appropriate than exact label matching. The CLAP baseline provides a reproducible estimate of intent alignment using off-the-shelf audio–text representations, while BM25 captures the contribution of textual descriptions alone, establishing complementary reference points for future retrieval and reranking methods. MER benchmarks reinforce this interpretation. OvR classifiers trained on CLAP embeddings achieve moderate micro-F1, yet exact-match accuracy remains low while Top-3 overlap is higher, reflecting the same graded structure observed in retrieval [21].

Several limitations should be noted. CalmSet is modest in scale (432 tracks), making it better suited for benchmarking, analysis, and fine-tuning than for training large models from scratch. CalmSet benchmarks *perceived* affective and therapeutic intent; annotations reflect crowd workers' emotional judgements of music purposefully composed for children with ASD [18, 27], not clinically validated therapeutic outcomes. Establishing such validation would require structured trials with ASD specialists, caregivers, and children. Annotations are derived from pre-qualified crowd workers rather than clinical experts and should therefore be interpreted as perceived affective or intent-related relevance rather than validated therapeutic outcomes. Emotion labels and descriptions are culturally and linguistically situated; differences in cultural interpretation of emotion and language use may introduce systematic bias into the annotations. GPT-generated descriptions may further reflect language priors that shape how annotators interpret intents. Our aggregation procedure prioritizes majority agreement, which improves consistency but may obscure minority interpretations or culturally specific readings of affect. Also, the retrieval benchmark includes only eight queries, limiting the robustness of statistical significance estimates. While this work reports initial retrieval and classification baselines, more advanced dense retrieval architectures and reranking strategies are not explored.

These limitations also point to directions for future work. Extensions include releasing standardized query sets and qrels-style relevance files, evaluating dense and hybrid retrieval pipelines, and reporting robustness analyses across multiple seeds or query resamples. Expanding annotations to continuous affective dimensions (e.g., valence–arousal) would enable regression-based evaluation and finer-grained relevance modeling. Bridging CalmSet toward clinical utility represents a longer-term but meaningful research direction. Concretely, retrieval models trained on CalmSet could be integrated with caregiver-facing interfaces, such as the uCue system from which the corpus originates [18], to surface music candidates that are then filtered or approved by therapists in a human-in-the-loop pipeline. Longitudinal studies tracking child behavioral and physiological responses to retrieved tracks could progressively ground affective relevance labels in therapeutic outcomes, enabling future versions of CalmSet to move closer to clinically validated benchmarking. Finally, cross-domain generalization studies—training on large-scale web music corpora and evaluating on CalmSet, and vice versa—would help characterize domain shift and robustness in affect-aware music retrieval.

7 Conclusion

We introduced CalmSet, a therapeutically inspired dataset of modular music designed for affect-aware music retrieval and music emotion recognition. The dataset combines CLAP-based audio representations, automatically generated textual descriptions, and crowdsourced relevance judgments within a transparent and reproducible annotation pipeline, enabling graded-relevance evaluation. Our contributions are threefold. First, we release a curated collection of modular music tracks annotated with ranked, multi-label affective intent signals that reflect how musical pieces may be interpreted and used in practice. Second, we establish baseline results for both emotion-conditioned music retrieval and multi-label music emotion recognition, demonstrating that while current audio–text representations capture coarse affective structure, substantial performance gaps remain—particularly for nuanced and context-dependent intents. Third, we position CalmSet as a reusable benchmark for studying graded relevance, intent ambiguity, and evaluation methodology in affect-sensitive music access, complementing existing MIR datasets that rely on categorical or binary labels. These findings position CalmSet as a challenging benchmark for studying subjective relevance, intent ambiguity, and evaluation methodology in music retrieval.

Acknowledgments

This work was supported by the National Science Foundation via the Accelerating Research Translation Program, Grant No. 2331440, and the University of Delaware's (UD) Institute for Engineering-Driven Health. Additional summer research support for Liam Stapley was provided by UD's Undergraduate Research Program. We also acknowledge the peoples of the Woi Wurrung and Boon Wurrung language groups of the eastern Kulin Nation on whose unceded lands ACM SIGIR 2026 was hosted. We pay our respects to their Elders past and present, and extend that respect to all Aboriginal and Torres Strait Islander peoples today and their continuing connection to land, sea, sky, and community.

References

[1] Christopher Akiki and Manuel Burghardt. 2021. MuSe: The Musical Sentiment Dataset. https://doi.org/10.34740/KAGGLE/DSV/2250730

[2] Anna Aljanaki, Frans Wiering, and Remco C Veltkamp. 2016. Studying emotion induced by music through a crowdsourcing game. *Information Processing & Management* 52, 1 (2016), 115–128.

[3] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. 2017. Developing a benchmark for emotional analysis of music. *PloS one* 12, 3 (2017), e0173392.

[4] Grace T Baranek, Fabian J David, Michele D Poe, Wendy L Stone, and Linda R Watson. 2006. Sensory Experiences Questionnaire: discriminating sensory features in young children with autism, developmental delays, and typical development. *Journal of child Psychology and Psychiatry* 47, 6 (2006), 591–601.

[5] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. 2011. The Million Song Dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*.

[6] Andreas F Ehmann[1]. 2008. The 2007 MIREX audio mood classification task: Lessons learned. In *ISMIR 2008: Proceedings of the 9th International Conference of Music Information Retrieval*. Lulu. com, 462.

[7] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[8] Peter Emerson. 2013. The original Borda count and partial voting. *Social Choice and Welfare* 40, 2 (2013), 353–358.

[9] Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*. 571–575. https://doi.org/10.21437/Interspeech.2021-698

[10] James Hiller and Susan C Gardstrom. 2018. The selection of music experiences in music therapy. *Music Therapy Perspectives* 36, 1 (2018), 79–86.

[11] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. 2022. Mulan: A joint embedding of music audio and natural language. *arXiv preprint arXiv:2208.12415* (2022).

[12] Hsiao-Tzu Hung, Joann Ching, Seungheon Doh, Nabin Kim, Juhan Nam, and Yi-Hsuan Yang. 2021. EMOPIA: A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-based Music Generation. In *Proc. Int. Society for Music Information Retrieval Conf.*

[13] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002).

[14] Rie Kamikubo, Kyungjun Lee, and Hernisa Kacorri. 2023. Contributing to accessibility datasets: Reflections on sharing study data by blind people. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[15] Hong Jin Kang, Fabrice Harel-Canada, Muhammad Ali Gulzar, Violet Peng, and Miryung Kim. 2024. Human-in-the-Loop Synthetic Text Data Inspection with Provenance Tracking. arXiv:2404.18881 [cs.HC] https://arxiv.org/abs/2404.18881

[16] Jaeyong Kang and Dorien Herremans. 2025. Are We There Yet? A Brief Survey of Music Emotion Prediction Datasets, Models and Outstanding Challenges. arXiv:2406.08809 [cs.SD] https://arxiv.org/abs/2406.08809

[17] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6769–6781.

[18] Abhishek Karwankar, Elise Ruggiero, Zoe Lipkin, Malika Karthik Iyer, Simon Brugel, Prerana Khatiwada, Daniel Stevens, and Matthew Louis Mauriello. 2025. *uCue: An Interactive Musical Interface to Enhance Formative Listening Experiences for Children with ASD*. Association for Computing Machinery, New York, NY, USA, 340–357. https://doi.org/10.1145/3713043.3727053

[19] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 39–48.

[20] Youngmoo E Kim, Erik M Schmidt, and Lloyd Emelle. 2008. Moodswings: A collaborative game for music mood label collection.. In *Ismir*, Vol. 8. Philadelphia, PA, 231–236.

[21] Kai R Larsen and Chih How Bong. 2016. A tool for addressing construct identity in literature reviews and meta-analyses. *Mis Quarterly* 40, 3 (2016), 529–552.

[22] Harin Lee, Elif Çelen, Peter Harrison, Manuel Anglada-Tort, Pol van Rijn, Minsu Park, Marc Schönwiesner, and Nori Jacoby. 2025. GlobalMood: A cross-cultural benchmark for music emotion recognition. arXiv:2505.09539 [cs.IR] https://arxiv.org/abs/2505.09539

[23] Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval* 3, 3 (2009), 225–331.

[24] Rashini Liyanarachchi, Aditya Joshi, and Erik Meijering. 2025. A Survey on Multimodal Music Emotion Recognition. *arXiv preprint arXiv:2504.18799* (2025).

[25] Athanasios Lykartsis, Andreas Pysiewicz, Henrik von Coler, and Steffen Lepa. 2013. The emotionality of sonic events: testing the geneva emotional music scale (GEMS) for popular and electroacoustic music. *Music & Emotion (ICME3)* (2013).

[26] Elysa J Marco, Leighton BN Hinkley, Susanna S Hill, and Srikantan S Nagarajan. 2011. Sensory processing in autism: a review of neurophysiologic findings. *Pediatric research* 69, 8 (2011), 48–54.

[27] Matthew Mauriello, Daniel Stevens, Elise Ruggiero, and Abhishek Karwankar. 2025. Interactive musical interface to enhance formative listening experiences for children with autism spectrum disorder. US Patent App. 19/201,395.

[28] Kimberly Sena Moore. 2013. A systematic review on the neural effects of music on emotion regulation: Implications for music therapy practice. *Journal of music therapy* 50, 3 (2013), 198–242.

[29] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Rangan Majumder Tiwary, R Majumder, and Li Deng. 2016. MS MARCO: A Human Generated Machine Reading Comprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches (NIPS 2016)*.

[30] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. 337–343.

[31] Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva. 2018. Musical texture and expressivity features for music emotion recognition. In *19th International Society for Music Information Retrieval Conference (ISMIR 2018)*. 383–391.

[32] Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology* 17, 3 (2005), 715–734.

[33] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* 3, 4 (2009), 333–389.

[34] Tetsuya Sakai and Ruihua Song. 2011. Evaluating diversified search results using per-intent graded relevance. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 1043–1052.

[35] Igor André Pegoraro Santana, Fabio Pinhelli, Juliano Donini, Leonardo Catharin, Rafael Biazus Mangolin, Valéria Delisandra Feltrim, Marcos Aurélio Domingues, et al. [n.d.]. Music4all: A new music database and its applications. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*.

[36] Roseann C Schaaf, Susan Toth-Cohen, Stephanie L Johnson, Gina Outten, and Teal W Benevides. 2011. The everyday routines of families of children with autism: Examining the impact of sensory processing difficulties on the family. *autism* 15, 3 (2011), 373–389.

[37] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. 2018. Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval* 7, 2 (2018), 95–116.

[38] Mark D Smucker, James Allan, and Ben Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. 623–632.

[39] Mark D Smucker, James Allan, and Ben Carterette. 2009. Agreement among statistical significance tests for information retrieval evaluation at varying sample sizes. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 630–631.

[40] Jacquelin A Speck, Erik M Schmidt, Brandon G Morton, and Youngmoo E Kim. 2011. A Comparative Study of Collaborative vs. Traditional Musical Mood Annotation.. In *ISMIR*, Vol. 104. 549–554.

[41] Lynda Tamine and Lorraine Goeuriot. 2021. Semantic information retrieval on medical texts: Research challenges, survey, and open issues. *ACM Computing Surveys (CSUR)* 54, 7 (2021), 1–38.

[42] Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2894–2913.

[43] Martin Theobald, Holger Bast, Debapriyo Majumdar, Ralf Schenkel, and Gerhard Weikum. 2008. TopX: efficient and versatile top-k query processing for semistructured data. *The VLDB Journal* 17, 1 (2008), 81–115.

[44] Sida Tian, Can Zhang, Wei Yuan, Wei Tan, and Wenjie Zhu. 2025. XMusic: Towards a Generalized and Controllable Symbolic Music Generation Framework. arXiv:2501.08809 [cs.SD] https://arxiv.org/abs/2501.08809

[45] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. 2007. 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*. 439–446.

[46] Hamed Valizadegan, Rong Jin, Ruofei Zhang, and Jianchang Mao. 2009. Learning to rank by optimizing ndcg measure. *Advances in neural information processing systems* 22 (2009).

[47] Ellen M Voorhees. 2003. Evaluating the evaluation: A case study using the TREC 2002 question answering track. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. 260–267.

[48] Ellen M. Voorhees and Chris Buckley. 2002. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Tampere, Finland) *(SIGIR '02)*. Association for Computing Machinery, New York, NY, USA, 316–323. https://doi.org/10.1145/564376.564432

[49] Ellen M Voorhees, Ian Soboroff, and Jimmy Lin. 2022. Can old TREC collections reliably evaluate modern neural retrieval models? *arXiv preprint arXiv:2201.11086* (2022).

[50] Shuo-Yang Wang, Ju-Chiang Wang, Yi-Hsuan Yang, and Hsin-Min Wang. 2014. Towards time-varying music auto-tagging based on CAL500 expansion. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[51] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels. Association for Computing Machinery, New York, NY, USA.

[52] Kate E Williams. 2018. Moving to the beat: Using music, rhythm, and movement to enhance self-regulation in early childhood classrooms. *International Journal of Early Childhood* 50, 1 (2018), 85–100.

[53] Huang Xie and Tuomas Virtanen. 2021. Zero-shot audio classification via semantic embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 1233–1242.

[54] Yi-Hsuan Yang and Homer H Chen. 2012. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3, 3 (2012), 1–30.

[55] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H. Chen. 2008. A Regression Approach to Music Emotion Recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 16, 2 (2008).

[56] Kejun Zhang, Hui Zhang, Simeng Li, Changyuan Yang, and Lingyun Sun. 2018. The PMEmo Dataset for Music Emotion Recognition. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval* (Yokohama, Japan) *(ICMR '18)*. ACM, New York, NY, USA, 135–142. https://doi.org/10.1145/3206025.3206037