

Euphrasis: An Interactive System for Music Preference Verbalization

Arisa Okamura¹, Kento Watanabe², Takayuki Nakatsuka², Kosetsu Tsukuda²,
Tian Cheng², Tomoyasu Nakano², and Masataka Goto²

¹ The Lawrenceville School, Lawrenceville, NJ 08648, USA
alicelavander@gmail.com

² National Institute of Advanced Industrial Science and Technology (AIST)
{kento.watanabe,takayuki.nakatsuka,k.tsukuda,tian.cheng,
t.nakano,m.goto}@aist.go.jp

Abstract. The music landscape today is becoming increasingly complex, with a growing diversity of structures and styles. This poses a challenge for many listeners, particularly those without formal training, in expressing their intuitive or abstract musical impressions and preferences in precise and descriptive language. Although current music captioning and tagging models can provide track-level descriptions, they offer limited cross-track insights and cannot fully support listeners to explore the underlying reasons for their music preferences. These limitations motivate exploring whether an interactive interface can strengthen listeners' ability to articulate the specific musical elements that underpin their taste. We propose Euphrasis, an interactive system that integrates music caption outputs from multiple tracks into a structured representation with a user-friendly interface. Euphrasis assists listeners in verbalizing emotional responses and abstract impressions by iteratively extracting and refining descriptors, specific and isolatable musical keywords, across multiple tracks. By comparing these descriptors, listeners recognize recurring patterns that clarify their music preferences. In preliminary user evaluations, participants produced more precise descriptors and reported higher self-awareness of taste, revealing the system's potential to enhance personal engagement with music.

Keywords: music preference verbalization · music captioning · augmented music-understanding interface.

1 Introduction

In an age where listeners devour millions of tracks spanning analog warmth, digital precision, and algorithmic novelty, few can explain why a song moves them. Beneath the surface of streaming lies a tangle of rhythmic motifs, textural subtleties, and micro-genre cues that remain largely invisible to the ear—and even more elusive to language [3]. Efforts in music captioning and tagging have contributed toward bridging this gap and facilitating the explication of music, as models developed for these tasks can generate natural language descriptions of individual tracks [6, 13, 17]. However, these models usually operate on isolated

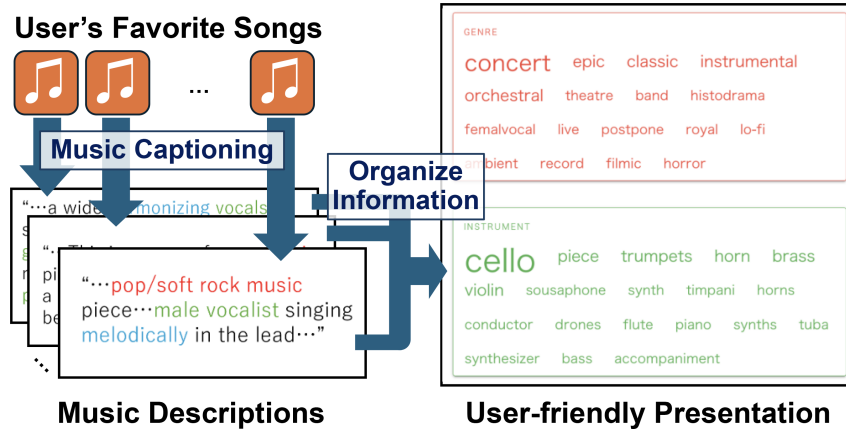


Fig. 1. A conceptual overview of Euphrasis. By systematically organizing and filtering the concrete details about individual music tracks generated by music captioning models, we aim to derive higher-level, abstract insights into each listener’s music preferences.

music tracks and thus lack mechanisms for cross-track comparison, provide little guidance on how different music pieces connect or how underlying musical features collectively shape a listener’s broader music preferences. This gap demonstrates the need for tools that not only describe music in isolation but also help listeners verbalize why they prefer certain styles, structures, or sonic elements over others.

Proposals for “augmented music-understanding interfaces” highlight the importance of improving listeners’ understanding of music [8, 9]. A structured way to verbalize the specific musical elements that contribute to their music preferences could deepen their self-awareness and comprehension of music in general. Although user-centered approaches in Music Information Retrieval (MIR) have gained traction [16, 21], tools that actively support listeners in grasping the intricacies of today’s rapidly expanding musical landscape remain relatively new. To suggest a new direction, we examine how an interactive, descriptor-driven interface can sharpen listeners’ articulation of why they enjoy particular music.

In this paper, we introduce Euphrasis, an interactive system that helps listeners articulate their music preferences by combining machine-generated captions with user feedback as shown in Figure 1. Euphrasis uses a music captioning model and Large Language Models (LLMs) to extract descriptors—keywords that describe isolatable attributes of music—from user-evaluated music tracks and classify them by category, supporting listeners who struggle to verbalize ‘why’ they like a track through iterative evaluation. We present a user-centered evaluation suggesting Euphrasis’s potential to develop listeners’ music understanding and support the realization of aspects in their music preferences, contributing to an emerging vision of augmented music-understanding interfaces.

2 Related Work

2.1 Recommendation Techniques

Numerous studies have explored ways of providing explanations for recommended items [1]. By reviewing the rationale attached to each recommendation, users can indirectly grasp their own preferences. More directly, some approaches generate textual summaries of a user’s tastes from past consumption histories and present these alongside recommended results [2]. However, in most cases, these explanations serve only as supplementary information. Because users cannot intervene in the system’s inference process, even if the explanations fail to reflect their self-perception, it is difficult to gain a deeper understanding of personal preferences from these explanations alone.

2.2 Music Captioning Techniques

Music captioning research aims to convert raw audio signals into succinct natural language descriptions by detecting attributes such as genre, mood, instrumentation, and overall structure [13]. Systems like MusCaps [17] and MusiLingo [4] adopt deep learning architectures that combine convolutional, recurrent, or transformer-based models to capture both temporal and spectral characteristics. Furthermore, LP-MusicCaps [5] addresses data scarcity by using LLMs to generate pseudo-captions. While recent advances show promise in bridging quantitative signal processing with human-interpretable language, most captioning models operate on individual tracks and lack comparative analysis. As a result, they cannot contextualize similarities or differences across songs or relate them to the listener’s broader musical preferences.

2.3 Joint Audio-Text Embedding

Joint audio-text embedding methods map music excerpts and textual phrases into a shared space using cross-modal contrastive learning [7, 12, 24, 25]. Although they excel at zero-shot retrieval, each track is still compressed into a single dense vector, describing *what* the music sounds like but not *why* a particular listener might enjoy it. Because this vector cannot isolate, weight, or discard specific musical facets, such embeddings remain opaque for preference modeling. Euphrasis therefore targets a higher level of abstraction, transforming content-level similarity into an interpretable model of music taste.

2.4 Interfaces on Music Understanding

Interactive interfaces offer visual and analytical tools for exploring music [10]. For instance, Instrudiver [22] is an interface that helps listeners explore a song’s instrumentation and arrangement through interactive elements like multi-colored pie charts and timeline graphs. Songle [11], on the other hand, visualizes a song’s

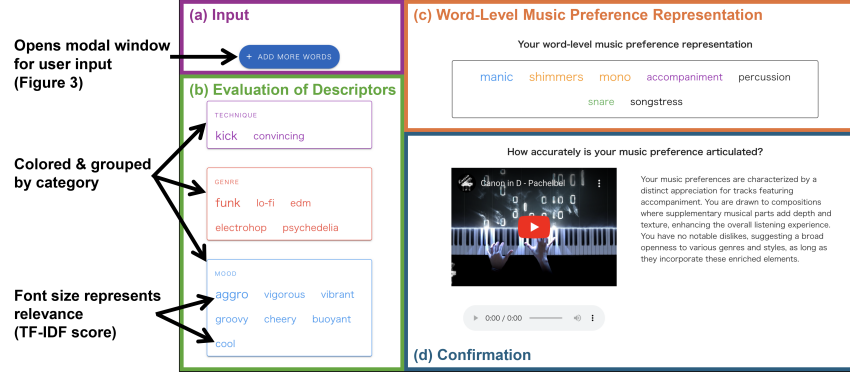


Fig. 2. Overview of Euphrasis, divided into four steps labeled (a) through (d). In step (a), a button opens a modal window (Figure 3) prompting user input about their music preferences. Step (b) displays potential musical descriptors related to these preferences. Step (c) serves as the result, collating all descriptors that the user confirmed describe their music preferences. Finally, step (d) provides synthesized insights, encouraging users to reflect and evaluate how accurately the system has captured their preferences.

timeline, chord progression, and melody, presenting analytical concepts in accessible formats that could lower the barrier to formal music training. Although both systems effectively illuminate critical aspects of music like instrumentation and structure, they largely emphasize objective analysis rather than guiding listeners to articulate the subjective qualities defining personal preferences; while they demystify certain technical details, they do not provide avenues for listeners to explain why specific attributes align with them on a more intrinsic level. These approaches offer objective visualization of music content but lack holistic interpretability, limiting listeners' ability to articulate preferences or reach essential musical understanding.

These limitations motivate a novel approach supporting both detailed and generalized exploration of preferences. Euphrasis offers an interactive feedback loop to encourage active verbalization and meaningful engagement, compiling objective analysis of music into comprehensive musical understanding.

3 Euphrasis

3.1 Interface Design

Euphrasis is an interactive system that converts users' subjective musical impressions into concrete descriptors. We define a musical descriptor as a short lexical unit—single word or fixed phrase—that denotes a specific, isolatable attribute of a recording (e.g., “syncopation,” “cello,” “lo-fi”). By systematically capturing and visualizing user feedback, it enables users to identify and articulate the elements

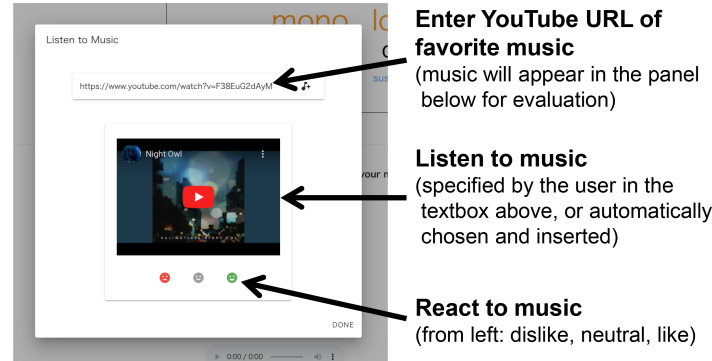


Fig. 3. Modal window triggered by button in Figure 2(a). The user listens to either a recommended music track or a track specified by entering a YouTube URL in the textbox at the top of the panel, and then rates the track as dislike, neutral, or like.

(e.g., instrumentation, genre, mood) that shape their music preferences. The workflow proceeds in four steps, labeled (a) through (d) in Figure 2:

(a) Input. To identify music that aligns with the user’s initial preferences, Euphrasis begins by collecting information about music the user likes and dislikes. The button labeled “+ADD MORE WORDS” in Figure 2(a) opens a modal window (Figure 3) where music is presented via a YouTube³ video. Users listen and respond with “like,” “neutral,” or “dislike” using the buttons at the bottom.

For music tracks rated positively or negatively, Euphrasis first invokes LP-MusicCaps [5] to generate concise text descriptions that touch upon the qualitative aspects like the mood and sound characteristics as well as the quantitative information like the tempo and instrumentation. A large language model (LLM), GPT-3.5-turbo [19], then extracts musically relevant keywords (hereafter referred to as descriptors) from these descriptions. The system stores these descriptors in separate lists based on whether the associated track is positively or negatively rated; each time a descriptor appears again, its frequency count is incremented, forming a dictionary-like data structure of descriptors and their frequencies. This structure enables Euphrasis to track musical traits that resonate—or fail to resonate—with individual users.

Users can directly specify a YouTube video through the textbox at the top of the modal window. Additionally, once users rate a music track, another LLM, GPT-4o [18], automatically provides further recommendations by inserting YouTube videos at the center of the modal window based on the list of descriptors from positively rated music. This allows users to continue rating music without having to search for additional tracks manually.

(b) Evaluation of Descriptors. Euphrasis next classifies the descriptors from step (a) into six color-coded categories using GPT-4o, based on labels most

³ YouTube is an online video-sharing platform (<https://www.youtube.com>).

Table 1. Six categories of music descriptors, assigned colors, and example descriptors.

category	color	example descriptors
genre	red	classical, lo-fi, edm
instrument	green	drum, keyboard, harp
mood	blue	groovy, vigorous, eerie
sound	orange	mono, midrange, distortion
technique	purple	arpeggio, kick, repeating melody
others	black	1980s, songstress, digital

frequently suggested in pilot experiments where a wide range of musics were provided into the system to see what categories would be generated. This step structures the user’s music preferences for easier analysis and interpretation. Music tracks from the liked and disliked sets populate two contrasting descriptor pools, but since pilot runs revealed that adding negatively weighted descriptors nearly doubled the decision load and overwhelmed users, only descriptors from positively reacted music are presented to the user for evaluation. The six categories, their assigned color, and example descriptors are shown in Table 1.

To highlight the descriptors that most effectively distinguish liked from disliked music tracks, Euphrasis applies a custom TF-IDF[20]-based weighting strategy. This method compares each descriptor’s frequency in the liked or disliked set of music tracks against its total occurrence. Specifically, the liked and disliked sets each serve as separate “documents,” where term frequency (TF) reflects how often a descriptor appears, and inverse document frequency (IDF) penalizes those appearing in both. Descriptors strongly tied to one sentiment receive higher weights, while common ones are downweighted.

These scores are reflected in the interface through a threshold that filters out low-scoring descriptors and by font size adjustments. Words with extremely low scores are recognized as noise and removed from the interface, while higher-scoring descriptors appear in larger font, drawing attention and indicating greater significance and relevance to the user. For example, in the “genre” category of Figure 2(b), “funk” has a higher TF-IDF score than “lo-fi,” “edm,” and other descriptors, thus resulting in a larger font. This helps reduce confusion and prevents cognitive overload from excessive or unfamiliar terms.

Descriptors are displayed in an augmented grid view, with color and font size reflecting their category and TF-IDF score. Users can confirm or remove each descriptor to articulate their preferences more precisely. Confirmed descriptors move to a summary box in the upper-right corner for clarity (Figure 2(c)).

Clicking a descriptor reveals a concise definition generated by GPT-4o [18], as well as links to the source track(s). This functionality bridges the gap between technical labels and users’ intuitive musical impressions. By iteratively adjusting descriptors over time, users gradually refine their preference profile and gain substantive insight into what incites strong responses to certain music.

(c) Word-Level Music Preference Representation. When users confirm that a music descriptor indeed describes their music preference accurately, they

can “save” the descriptor as one of the components of their word-level music preference representation in step (c). The compiled words represent the final result, which the users can consult to understand why certain recommendations emerge. This display retains the color-coding of categories (e.g., instrumentation, style, mood) and applies the same TF-IDF-based weighting. By consolidating terms that the user explicitly validates, Euphrasis ensures the final profile accurately reflects true preferences rather than irrelevant words. This structured snapshot explicitly summarizes user preferences through specific descriptors like “manic (music),” “mono”, and “snare” rather than vague and interpretable adjectives such as “mysterious.” Knowing these vocabularies lets users recognize and communicate those cues in new music; conversely, users can collapse terms to category level (e.g., Instrument→Strings) when they prefer coarser summaries. By making descriptors explicit, Euphrasis provides transparency and control typically absent in recommendation engines, clarifying both what resonates with users and why.

(d) Confirmation. In the final step, users verify the accuracy of their refined preference profile. As shown in Figure 2(d), Euphrasis presents a music recommendation along with a synthesized description of the user’s music preferences, both generated by GPT-4o based on all collected data, including descriptors and their frequency counts from both descriptor sets, as well as the set of confirmed words. This layout allows users to assess whether their validated descriptors accurately match the recommended track and to explore further options aligned with these descriptors. If suggestions appear off-target, users can revert to step (b) to adjust their chosen descriptors. Through this iterative process, Euphrasis converges on a word-level portrayal of each user’s musical identity, ultimately transforming subjective impressions into concrete, precise insights.

3.2 Implementation

Euphrasis is implemented as a web-based application, making it accessible to diverse users regardless of their level of expertise in music analysis. The backend integrates LP-MusicCaps [5] to convert audio into textual captions, while GPT-3.5-turbo [19] extracts relevant musical descriptors with a prompt of the form:

“Respond in JSON, with key ‘len’ and ‘keywords’. ‘len’ is an integer that shows how many words the ‘keywords’ list contains. ‘keywords’ is a list of words or phrases, which are musical terms & any related terminology from the following description: [original result]. Keep out any general concept terms like ‘tempo’ and words that talk about quality, but add other words that are as technical as possible & unique to the specific music. Additionally, predict the genre of the music from the description and add to the ‘keywords’ list.”

For example, LP-MusicCaps [5] might generate a caption like “melancholic strings with prominent piano melody and orchestral textures” for the music track that a user rates positively. GPT-3.5-turbo then extracts discrete descriptors (e.g., “melancholic,” “string,” “piano melody,” “orchestral”), collating the ‘keywords’ list that can be analyzed and evaluated by the user in step (b). Subsequently,

GPT-4o [18] receives a list of ‘keywords’ extracted from positively reacted music and adds music recommendation for the input modal window in step (a):

“Respond in json, with key ‘len’ and ‘musics’. ‘len’ is an integer that shows how many musics the ‘musics’ list contain. Please suggest multiple songs into the ‘musics’ list, in the format of title + artist, from the following list of words: [list of descriptors from positively reacted music].”

Then, GPT-3.5-turbo classifies each descriptor by the following prompt:

“Respond in one word without period. What is the category of the word [descriptor]? choose from the following: genre, instrument, mood, sound, technique, others.” Definitions and contexts for each descriptor are subsequently generated via: “Concisely explain the meaning of the word ‘[word]’ in the context of music & song characteristics.”

Lastly, the music preference description in step (d) is produced by GPT-4o:

“Respond in second person. Write a short & concise paragraph that best describes the music preference & music taste of an audience with the following information. Dictionary positiveWords is a list of words the user preferred as a key and its frequency/weight as the corresponding value. Dictionary negativeWords is also a list of words the user disliked as a key and its frequency/weight as the corresponding value. positiveWords: [dictionary of descriptors from positively reacted music including the confirmed descriptors], negativeWords: [dictionary of descriptors from negatively reacted music].”

All LLM calls were made through the OpenAI API. We selected GPT-4o as the default because, at the time of development, it provided the best balance of latency and accuracy among available models. For steps that involved either very long prompts or large batch outputs (e.g., bulk keyword extraction from captions, step (a)), we switched to the less resource-intensive GPT-3.5-turbo to control computation cost without compromising downstream quality.

4 Evaluation

To investigate the effectiveness of Euphrasis, we conducted two exploratory assessments: a small-scale test showing how the system distinguishes users with contrasting music preferences, and a qualitative study focused on how the system influences users’ understanding and ability to articulate their preferences.

4.1 Preliminary Observations

We performed an informal check using two hypothetical user profiles to represent contrasting musical backgrounds. One profile reflected a preference for dark, ominous, and modern pop music with strong percussive elements, while the other simulated a preference for classical music rich in orchestral instrumentation and

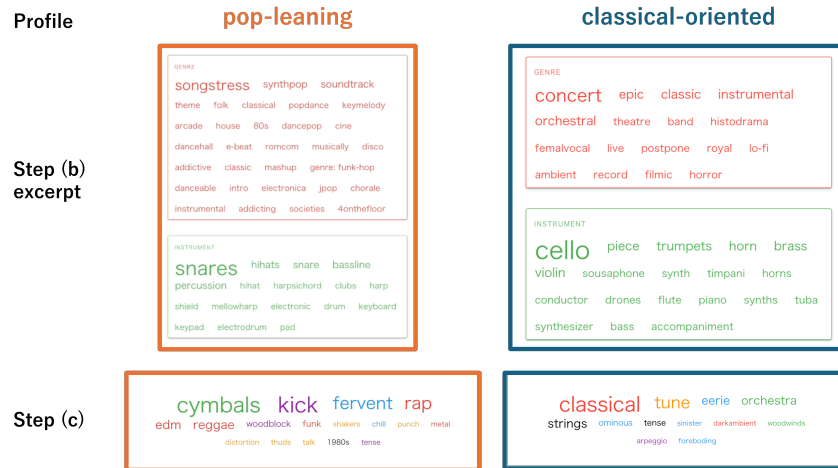


Fig. 4. Example snapshots of two hypothetical profiles (step (b) and step (c) from Figure 2). Step (b) panels show only cropped regions from larger intermediate outputs. Step (c) panels were captured in a separate trial of the same experiment. The left-hand side corresponds to a pop-leaning profile, and the right-hand side corresponds to a classical-oriented profile.

strings. Both users provided sufficient feedback on a curated set of music tracks, allowing Euphrasis to generate a word-level representation of their preferences.

As shown in Figure 4, in step (b), the pop-leaning profile on the left-hand side yielded descriptors expressing some characteristics of modern pop such as “songstress,” “percussion,” and “80s.” In step (c), the vocabulary was filtered further, and listed terms like “cymbals,” “kick,” and “reggae,” which refine and narrow the portrayal of the profile’s musical preference. In contrast, the classical-oriented profile on the right-hand side prominently features descriptors like “orchestral,” “cello,” and “instrumental” in step (b), and descriptors like “strings,” “arpeggio,” and “classical” in step (c), indicating a focus on orchestral textures and instrumentation. Comparing steps (b) and (c), the more concise set of terms in step (c) makes it easier for users to grasp the portrait of their musical preferences. Although this scenario was limited in scope, the differences between profiles consistently manifested across both steps, suggesting that Euphrasis can indeed surface distinct musical attributes tailored to divergent preferences. Beyond instrumentation, the system also identified mood- and technique-related descriptors, highlighting its broader potential for revealing how stylistic factors shape individual listening preferences.

4.2 User Study

To provide a more detailed examination of Euphrasis, we conducted a user study with ten high school students (seven females and three males) aged 15 to 17,

recruited through convenience sampling by the first author, who was a high school student at the time of this study. The participants had diverse musical backgrounds and preferences: four participants primarily favored modern pop and electronic music, two expressed strong interests in classical and orchestral genres, one participant favored rock music from the 70s, and others had vague preferences described by adjectives such as “peaceful,” “soothing,” and “energetic.”

The experiment was conducted following the steps:

1. Initial Survey (3-5 minutes). Participants received a brief overview of Euphrasis, outlining the purpose of the study and instructions on system usage. Participants wrote a brief paragraph describing their general music preferences and favorite genres without any system assistance.
2. Interaction with Euphrasis (15-20 minutes). Participants started interacting with Euphrasis by specifying one music they frequently listen to, and explored the system for 15-20 minutes after a brief explanation of the system’s overview and features. Participants were explicitly instructed to actively explore suggested descriptors and observe how selections influenced subsequent recommendations.
3. Final Survey (5 minutes). After the interaction, participants completed a questionnaire that included quantitative and qualitative assessments. Quantitative evaluations used a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree) to assess the clarity, usefulness, and impact of the system. Qualitative feedback involved open-ended questions asking participants to elaborate on new insights and experiences during the session.

Each participant individually interacted with the system in a quiet classroom environment after school hours to minimize external distractions.

The quantitative portion of the study comprised three questions:

- Q1** How much did the system help you develop your understanding/articulation of your music preference?
- Q2** How much did you develop a more precise and/or accurate verbalization of your music preference?
- Q3** How intuitive was the system?

Figure 5 shows the results. The median ratings were 4 for the first question, 4 for the second, and 5 for the third. These ratings suggest that Euphrasis effectively clarifies and communicates personal music preferences.

We also observed Euphrasis’ effectiveness through qualitative feedback. Nine participants reported uncovering new musical facets that they had not previously recognized about their music preferences. One participant remarked, “*I found that I don’t like overly-dominant sounds/percussions, especially really harsh drum beats,*” while another noted, “*I found new words like vivid, expressive, and string to be particularly interesting because it made me realize that a lot of my music preference does involve aspects like string incorporated into the rock that I didn’t recognize before.*” This process was often fueled by encountering previously unfamiliar descriptors. Interestingly, one participant reflected that the system

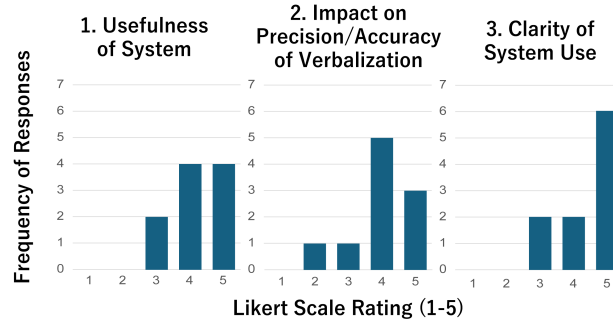


Fig. 5. Quantitative survey results (n=10).

provoked new discoveries without necessarily introducing new terminology, stating that they “*never really thought to apply [the music terms they already knew] to my own preference before.*”

Comparing the words that participants have listed as words that describe their music preferences, the results after the experiment presented more words that were specific, technical, and detailed. For example, a participant that listed “*kpop, afro beats, jazz, funky beats, pop catchy melody, synchopated rhythm, calm classical, lit*” as their descriptors listed the following list after the experiment: “*funky, female vocal, groovy, riff, tense, cheery, poignant, pluck, rnb, lento, ostinato, romantic, soft, synthesizer, bass, kicks, harmony, symbols, feelgood, edm, dance, soul*”.

Alongside these revelations, participants appreciated the system’s capacity to reveal pervasive themes in their broader listening habits. “*Reading the definitions of the words I realized that most of them corresponded to some part of my playlist, including songs that I did not put into the system,*” said one participant, underscoring how insights could extend beyond the specific music tracks utilized for analysis. While two participants indicated a desire for extended use to capture more details of their preferences, the rest agreed that Euphrasis substantially enhanced their initial understanding and articulation of music preferences. These findings underscore the system’s potential not only for immediate insight but also as a foundation for deeper, iterative exploration of music.

5 Conclusion

Euphrasis is an interactive system that translates listeners’ subjective musical impressions into precise, shareable descriptors by combining music captioning, LLMs, and TF-IDF weighting, enabling users to articulate vague preferences, uncover relevant terminology, and recognize meaningful listening patterns through iterative exploration. Quantitative and qualitative evaluations suggested that participants gained clearer insight into their musical affinities. At the same

time, limitations emerged: because the captioning model (LP-MusicCaps) is trained only on non-copyrighted audio, crucial descriptors for classical works and copyrighted modern tracks are often omitted. In addition, long-tail imbalances in training corpora bias both GPT-4o-driven retrieval and summarization toward Western, mainstream repertoires, risking the marginalization of underrepresented regions and genres [14, 15, 23]. Addressing these issues will require broadening training data across cultures and styles, fine-tuning LLMs with domain-specific corpora, considering more objective measures for user studies, and expanding the participant pool for user evaluations across age and cultural groups in order to further validate effectiveness and support inclusive, human-centered music discovery. By pursuing these directions, Euphrasis can not only improve its immediate interface but also generate a growing repository of user-validated descriptor-track pairs for future MIR research, ultimately enriching how diverse audiences engage with and appreciate music.

Acknowledgments. This work was supported in part by the Hutchins Science Scholars Program, the JST Science and Technology Challenge Program for Next Generation, and JST CREST Grant Number JPMJCR20D4, Japan.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Afchar, D., Melchiorre, A., Schedl, M., Hennequin, R., Epure, E., Moussallam, M.: Explainability in music recommender systems. *AI Magazine* **43**(2), 190–208 (2022)
2. Balog, K., Radlinski, F., Arakelyan, S.: Transparent, scrutable and explainable user models for personalized recommendation. In: *Proceedings of ACM SIGIR 2019*. pp. 265–274 (2019)
3. Conrad, F., Corey, J., Goldstein, S., Ostrow, J., Sadowsky, M.: Extreme re-listening: Songs people love... and continue to love. *Psychology of Music* **47**(2), 158–172 (2019)
4. Deng, Z., Ma, Y., Liu, Y., Guo, R., Zhang, G., Chen, W., Huang, W., Benetos, E.: MusiLingo: Bridging music and text with pre-trained language models for music captioning and query response. In: *Proceedings of NAACL 2024 Findings*. pp. 3643–3655 (2024)
5. Doh, S., Choi, K., Lee, J., Nam, J.: LP-MusicCaps: LLM-based pseudo music captioning. In: *Proceedings of ISMIR 2023*. pp. 409–416 (2023)
6. Drossos, K., Lipping, S., Virtanen, T.: Clotho: An audio captioning dataset. In: *Proceedings of IEEE ICASSP 2020*. pp. 736–740 (2020)
7. Elizalde, B., Deshmukh, S., Ismail, M.A., Wang, H.: CLAP: Learning audio concepts from natural language supervision. In: *Proceedings of IEEE ICASSP 2023*. pp. 1–5 (2023)
8. Goto, M.: Augmented music-understanding interfaces. In: *Proceedings of SMC 2009 Inspirational Session* (2009)
9. Goto, M.: Music listening in the future: Augmented music-understanding interfaces and crowd music listening. In: *Proceedings of the AES 42nd International Conference on Semantic Audio*. pp. 21–30 (2011)

10. Goto, M., Dannenberg, R.B.: Music interfaces based on automatic music signal analysis: new ways to create and listen to music. *IEEE Signal Processing Magazine* **36**(1), 74–81 (2018)
11. Goto, M., Yoshii, K., Fujihara, H., Mauch, M., Nakano, T.: Songle: A web service for active music listening improved by user contributions. In: *Proceedings of ISMIR 2011*. pp. 311–316 (2011)
12. Guzhov, A., Raue, F., Hees, J., Dengel, A.: AudioCLIP: Extending CLIP to image, text and audio. In: *Proceedings of IEEE ICASSP 2022*. pp. 976–980 (2022)
13. Kim, C.D., Kim, B., Lee, H., Kim, G.: AudioCaps: Generating captions for audios in the wild. In: *Proceedings of NAACL 2019*. pp. 119–132 (2019)
14. Kowald, D., Schedl, M., Lex, E.: The unfairness of popularity bias in music recommendation: A reproducibility study. In: *Proceedings of ECIR 2020*. p. 35–42 (2020)
15. Levy, M., Bosteels, K.: Music recommendation and the long tail. In: *Proceedings of WOMRAD 2010*. pp. 55–58 (2010)
16. Liem, C.C., Müller, M., Eck, D., Tzanetakis, G., Hanjalic, A.: The need for music information retrieval with user-centered and multimodal strategies. In: *Proceedings of ACM MM 2011 MIRUM Workshop*. pp. 1–6 (2011)
17. Manco, I., Benetos, E., Quinon, E., Fazekas, G.: MusCaps: Generating captions for music audio. In: *Proceedings of IJCNN 2021*. pp. 1–8 (2021)
18. OpenAI: Hello gpt-4o (2024), <https://openai.com/index/hello-gpt-4o/>
19. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Aspell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155* (2022)
20. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. *Information processing & management* **24**(5), 513–523 (1988)
21. Schedl, M., Knees, P., McFee, B., Bogdanov, D.: Music recommendation systems: Techniques, use cases, and challenges. In: *Recommender Systems Handbook*, pp. 927–971. Springer US (2021)
22. Takahashi, T., Fukayama, S., Goto, M.: Instrudiver: A music visualization system based on automatically recognized instrumentation. In: *Proceedings of ISMIR 2018*. pp. 561–568 (2018)
23. Tao, Y., Viberg, O., Baker, R.S., Kizilcec, R.F.: Cultural bias and cultural alignment of large language models. *PNAS Nexus* **3**(9), pgae346 (2024)
24. Wu, H.H., Seetharaman, P., Kumar, K., Bello, J.P.: Wav2CLIP: Learning robust audio representations from CLIP. In: *Proceedings of IEEE ICASSP 2022*. pp. 4563–4567 (2022)
25. Wu, S., Yu, D., Tan, X., Sun, M.: CLaMP: Contrastive language-music pre-training for cross-modal symbolic music information retrieval. In: *Proceedings of ISMIR 2023*. pp. 157–165 (2023)