EmoTun: Physiologically-Anchored AI Scaffolding for Therapeutic Musical Co-Creation between Humans and AI

Chan Qiu
Southern University of Science and Technology
Shenzhen, China
12433271@mail.sustech.edu.cn

Sirui Cheng
Southern University of Science and Technology
Shenzhen, China
12433266@mail.sustech.edu.cn

Tuo Zhang
Southern University of Science and Technology
Shenzhen, China
12433280@mail.sustech.edu.cn

Wenxin Yan
Southern University of Science and Technology
Shenzhen, China
12433279@mail.sustech.edu.cn

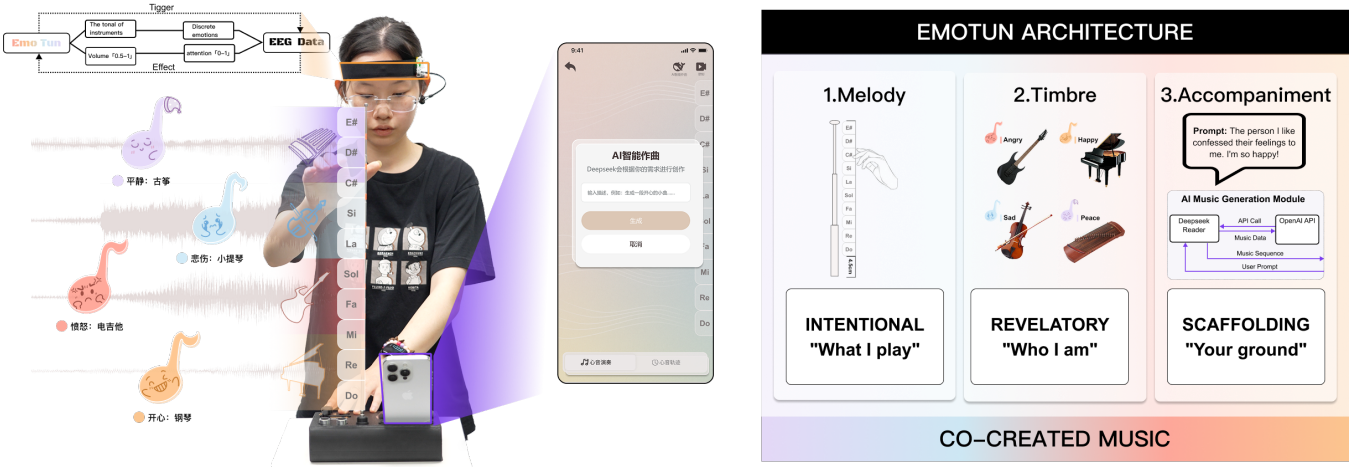

Figure 1: caption

Abstract

Musical improvisation benefits expressive arts therapy, yet traditional instruments exclude non-musicians, and AI music generation typically produces polished artifacts decoupled from the user's lived emotional experience. We present **EmoTun**, a three-layer musical co-creation system in which the user controls melody through hand gestures, EEG-driven emotion classification continuously modulates timbre, and AI-generated loop accompaniment provides harmonic scaffolding. We introduce *physiologically-anchored scaffolding*, in which AI support is grounded in the user's real-time embodied signals rather than arbitrary rules. Building this system surfaced three design tensions: the therapeutic ambiguity of physiological veridicality, the ethics of involuntary emotional disclosure, and the boundary between scaffolding and creative dependency. We ask whether emotional authenticity in human-AI co-creation constitutes a meaningful design objective or a seductive illusion.

DIS Companion '26, Singapore, Singapore

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

CCS Concepts

• **Human-centered computing** → **Interaction design process and methods**; **Accessibility systems and tools**; *Collaborative and social computing*; • **Applied computing** → *Sound and music computing*.

Keywords

physiologically-anchored scaffoldingmusical co-creationexpressive arts therapyemotional authenticity

ACM Reference Format:
Chan Qiu, Sirui Cheng, Tuo Zhang, and Wenxin Yan. 2026. EmoTun: Physiologically-Anchored AI Scaffolding for Therapeutic Musical Co-Creation between Humans and AI. In *Designing Interactive Systems Conference (DIS Companion '26), June 13–17, 2026, Singapore, Singapore.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Musical improvisation has emerged as a compelling modality for emotion regulation in expressive arts therapy. Prior work shows that improvisation engages three core dimensions of emotional processing—expression, awareness, and confrontation with emotional pain—which collectively predict therapeutic recovery from depression [5]. Process-level evidence further demonstrates that

improvisation induces measurable shifts across all five components of Scherer's Component Process Model (expression, feeling, bodily response, appraisal, and action tendency), with participants consistently reporting joy, social attunement, and physiological relaxation [1].

Yet this therapeutic potential remains largely unrealized due to a fundamental barrier to accessibility: traditional instruments require years of deliberate practice before users achieve expressive fluency, thereby systematically excluding those without formal musical training [8]. Recent advances in AI-powered music generation have sought to bridge this gap. Systems such as Amuse [3] enable collaborative songwriting through multimodal inputs, while commercial platforms like Suno and Udio promise to democratize composition altogether. However, critical analyses reveal that such democratization often functions as rhetoric rather than genuine empowerment [4]: when AI generates music *on behalf of* the user, the creative act becomes decoupled from the creator's lived experience. Under such conditions, users gain little opportunity for emotional self-awareness or expressive agency, undermining the very purpose of therapeutic musical engagement. This tension between accessibility and authenticity remains unresolved.

We argue that resolving this dual tension requires rethinking AI's role from *generator* to *scaffold*. Scaffolding, originally developed in educational psychology by Wood, Bruner, and Ross [7], refers to the temporary support a more capable partner provides to help a learner accomplish tasks just beyond their independent ability. In the context of musical co-creation, we reinterpret scaffolding as AI-provided musical context that enables non-musicians to engage in expressive improvisation while remaining subordinate to the user's own creative agency. We instantiate this idea in **EmoTun**, a three-layer musical co-creation system in which AI provides contextual loop accompaniment that supports rather than replaces the user's expressive acts. Critically, EmoTun introduces **Physiologically-Anchored Scaffolding**: timbre is continuously modulated by real-time EEG-based emotion classification, ensuring that the system's responsiveness remains coupled to the user's genuine affective state rather than arbitrary rules or learned preferences.

2 *EmoTun*: Toward Emotionally-Grounded Human-AI Musical Dialogue

EmoTun integrates three layers, melody, timbre, and accompaniment, into a real-time co-creation pipeline for therapeutic improvisation. Each layer occupies a distinct position within the physiologically anchored scaffolding framework: the user controls melody through hand gestures, EEG-driven emotion classification modulates timbre, and an AI language model provides contextual harmonic context.

2.1 System Architecture

The system architecture consists of three layers, each with a distinct role in the co-creation process (Figure 1):

Conscious Melodic Control. An ultrasonic sensor tracks the distance between the user's hand and the device, mapping distance to pitch in real time. This gives non-musicians an intuitive, gestural entry point into melodic expression without requiring knowledge of scales, notation, or finger placement.

Subconscious Timbre Modulation. A consumer-grade EEG module (TGAM) captures brainwave signals, which a Fast Fourier Transform (FFT) + K-Nearest Neighbors (KNN) classifier maps to one of four emotional states. Each state drives a different instrumental timbre: calm → guzheng, happy → piano, sad → violin, angry → electric guitar. The mapping is based on established acoustic correlates of emotion [2]—attack slope, spectral centroid, and spectral flux—so that the timbral shift is perceptually congruent with the affective state it represents.

AI-Generated Harmonic Scaffolding. A large language model (DeepSeek) receives short textual descriptions of the user's emotional context (derived from the EEG classification) and generates loop-based accompaniment in a matching musical style. The loop provides harmonic context without dictating the user's melody, functioning as a temporary support surface rather than a co-composer.

2.2 Technical Implementation

EmoTun is deployed on a Raspberry Pi 4 with a TGAM EEG module for cortical signal acquisition (512 Hz sampling rate), an HC-SR04 ultrasonic sensor for gesture-to-pitch distance mapping, and a Flask-based backend that uses WebSocket for real-time data flow between sensor inputs and audio output. Emotion classification employs FFT-based frequency-domain feature extraction, followed by a K-nearest neighbors classifier (k = 5) trained on four-class-labeled EEG data. The AI accompaniment layer submits emotion-derived textual prompts to the DeepSeek API, which returns MIDI loop accompaniments matched to the requested affective atmosphere. The full pipeline—from EEG acquisition through classification, timbre modulation, and loop synthesis—operates at sub-second latency, satisfying the temporal requirements of real-time musical improvisation in therapeutic contexts.

2.3 Physiologically-Anchored Scaffolding

We ground EmoTun's design in the concept of scaffolding, originally developed by Wood, Bruner, and Ross to describe how a more capable partner provides temporary support that enables a learner to operate at the edge of their competence [7]. Effective scaffolding exhibits two defining properties: it makes a task accessible that would otherwise exceed the learner's current ability, and it is designed to fade as competence grows, so that the learner ultimately achieves independence. In therapeutic contexts, however, this second property requires qualification: a client in expressive arts therapy may not be on a trajectory toward musical independence, and the presence of support can be therapeutically meaningful in itself—a form of being accompanied rather than a skill to be outgrown.

EmoTun's three-layer architecture—described in Section 2.1—translates scaffolding into a musical co-creation context while respecting this therapeutic nuance. *Layer 1 (Conscious Melodic Control)* instantiates scaffolding in its most literal sense: by collapsing the complex motor-symbolic mappings of traditional instruments into a single continuous gesture (distance → pitch), it reduces task demands, allowing non-musicians to produce intentional melodic phrases immediately. Critically, this layer reserves full voluntary control for the user, with no AI mediation. *Layer 2 (Subconscious Timbre Modulation)* addresses the authenticity dimension: rather than asking the user to select a timbre that matches their mood—a cognitive

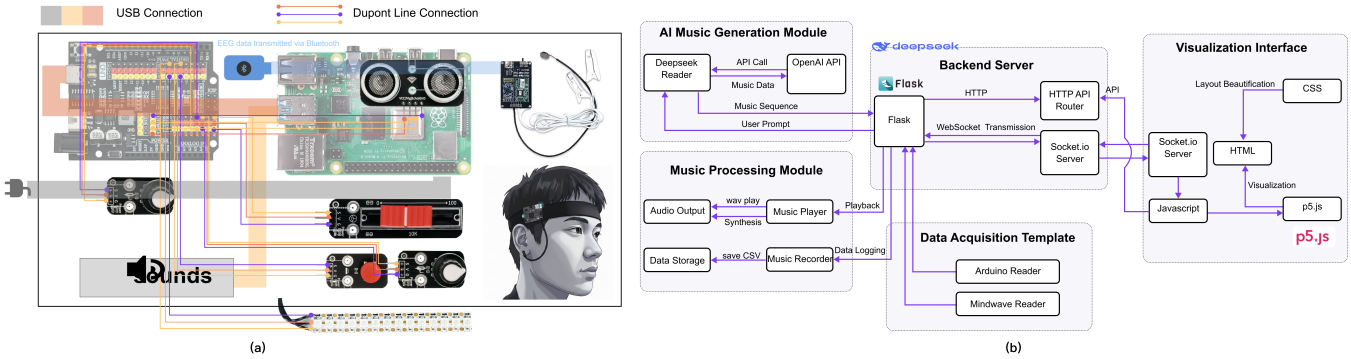

Figure 2: (a) Hardware connection diagram; (b) Software System Architecture Diagram

act that presumes the very emotional awareness therapy seeks to cultivate—the system derives timbre directly from the user's EEG, producing an acoustic reflection of embodied state that the user did not author but cannot disown. *Layer 3 (AI-Generated Harmonic Scaffolding)* is co-determined: it derives from the same EEG classification as timbre. Still, it is rendered through an LLM that interprets the emotional state into a stylistically appropriate harmonic context. Because the loop is generated from physiological rather than preference signals, it resists the drift toward statistical optimization and instead reflects who the user is in this moment.

We term this design principle physiologically-anchored scaffolding: AI-provided support grounded in the user's embodied affective state rather than in cognitive self-report or behavioral history.

3 Discussion

Building EmoTun surfaced three tensions that we believe are central to the design of physiologically-responsive AI systems for expressive arts therapy.

3.1 Physiological Honesty and the Empathy Illusion

EmoTun's EEG-to-timbre pipeline bypasses the user's cognitive filter: it does not ask the user to *report* how they feel, nor does it require them to *choose* a timbre that matches their mood. The system reads and responds. On one hand, this creates a form of emotional honesty that is rare in human-computer interaction—the system reflects what the body signals, not what the user is willing or able to articulate. On the other hand, this "honesty" comes with an uncomfortable corollary. The user cannot perform a socially acceptable emotion. If they are angry, the system plays an electric guitar regardless of whether anger feels safe or appropriate in the moment. The system, in effect, refuses to be lied to.

This raises a question about the nature of empathy in physiologically-driven systems. When EmoTun shifts from piano to violin because the EEG classifier detects sadness, is the user experiencing *being understood* or *being monitored*? Prior work on emotionally aware smart instruments has found that musicians sometimes trust AI's emotional reading more than their own self-assessment [6]—a finding that becomes ethically charged when transposed to therapeutic settings, where vulnerability is already heightened. The system's responsiveness may create an *illusion* of empathy—the AI "gets" me—while, in reality, it is a classification pipeline with no capacity for understanding. Whether this illusion is therapeutically productive (it creates a felt sense of co-presence) or therapeutically deceptive (it substitutes signal processing for genuine relational attunement) is, we believe, an open question for the field.

3.2 The Ethics of Involuntary Emotional Disclosure

EEG-based emotion classification introduces a privacy dynamic that has no analogue in self-report-based systems. When a user fills out a mood questionnaire, they curate their response: they decide what to disclose. When an EEG headband streams raw brainwave data to a real-time classifier, the user discloses emotional information they may not yet have consciously registered. The system knows before the user does.

This temporal asymmetry between classification and self-awareness creates a specific ethical challenge for therapeutic HCI. On the one hand, surfacing unrecognized emotions can be therapeutically valuable—it fosters awareness, a core mechanism of emotion regulation in music therapy [5]. On the other hand, the system may classify incorrectly (78% accuracy means roughly one in five classifications is wrong), and a misclassification in a therapeutic context is not a neutral error: telling a calm person they are angry, through the medium of an electric guitar timbre they did not choose, could be destabilizing. The question is not simply whether such systems should *display* their emotion classifications to the user—it is whether the very act of *responding* to an inferred emotion constitutes a form of feedback that the user cannot opt out of. What consent frameworks are needed when a physiological AI responds to emotions that the user may not yet know they are feeling?

3.3 Scaffolding Boundaries: When Does Support Become a Crutch?

EmoTun's AI loop accompaniment is deliberately passive: it provides harmonic context but does not intervene in melody or timbre. This is, in effect, the lowest rung of a broader ladder of AI initiative:

L1 Passive loop provider. The user retains full agency over melody and timbre; the AI merely supplies a harmonic bed. This is the current EmoTun configuration.

- **L2 Responsive style adapter.** The AI adjusts loop genre or rhythm in response to EEG trends, moderately reducing user burden. For example: "you seem more agitated; shifting to a calmer harmonic bed."
- **L3 Conversational improviser.** The AI initiates musical phrases and the user responds, functioning as an equal co-improviser in a therapeutic duet. User agency is lowest here.

Each step up this ladder reduces the user's creative burden—which is valuable for accessibility—but also reduces the user's creative agency. At what point does scaffolding become a crutch?

Wood et al.'s original scaffolding framework [7] emphasizes *fading*: the support structure should gradually withdraw as the learner's competence grows, so that the learner ultimately achieves independence. But therapeutic contexts complicate this logic. A client in expressive arts therapy may not be on a trajectory toward musical independence; the *presence* of support may be therapeutically meaningful in itself—a form of being accompanied rather than a skill to be outgrown. If the AI's loop accompaniment is slowly reduced as the user's melodic fluency improves, has the system empowered the user or quietly withdrawn the very relational quality that made the experience therapeutic? We do not have a resolution to this tension.

4 Conclusion

EmoTun is a modest system: a sensor, a classifier, a loop generator. But building it forced us to confront questions that extend far beyond our specific implementation. We found that anchoring AI support in physiological signals can indeed address the dual challenge of accessibility and authenticity—letting AI scaffold without replacing. But we also found that physiological anchoring introduces its own set of tensions: the uncomfortable honesty of an uncensored emotional signal, the privacy implications of responding to emotions before the user has named them, and the unresolved question of whether scaffolding should ever fade.

Acknowledgments

We would like to express our sincere gratitude to our supervisors, Mirna Zordan, Fang Wan, and Liang Hao, for their invaluable guidance and support throughout this research. We also extend our heartfelt thanks to our innovation course instructor, Yongsheng Ma, for his insightful advice and encouragement during the development of this project.

GenAI Usage Disclosure

The authors used ChatGPT and DeepL to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and assume full responsibility for the content of the publication.

References

[1] Sonja Aalbers, Annemieke Vink, Martina de Witte, Kim Pattiselanno, Marinus Spreen, and Susan van Hooren. 2021. Feasibility of Emotion-Regulating Improvisational Music Therapy for Young Adult Students with Depressive Symptoms: A Process Evaluation. *Nordic Journal of Music Therapy* 31, 2 (7 2021), 133–152. doi:10.1080/08098131.2021.1934088

[2] Patrik N. Juslin and Petri Laukka. 2003. Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code? *Psychological Bulletin* 129, 5 (2003), 770–814. doi:10.1037/0033-2909.129.5.770

[3] Yewon Kim, Sung-Ju Lee, and Chris Donahue. 2025. Amuse: Human-AI Collaborative Songwriting with Multimodal Inspirations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, 1–28. doi:10.1145/3706598.3713818

[4] Liam Pram and Fabio Morreale. 2025. Opening Musical Creativity? Embedded Ideologies in Generative-AI Music Systems. doi:10.48550/ARXIV.2508.08805

[5] Suvi Saarikallio, Petri Toiviainen, Olivier Brabant, Nerdinga Snape, and Jaakko Erkkilä. 2022. Music Therapeutic Emotional Processing (MEP): Expression, Awareness, and Pain Predict Therapeutic Outcome. *Psychology of Music* 51, 1 (4 2022), 140–158. doi:10.1177/03057356221087445

[6] Luca Turchet, Camille Baralon, and Cumhur Erkut. 2024. Musician-AI Partnership Mediated by Emotionally-Aware Smart Musical Instruments. *International Journal of Human-Computer Studies* 191 (2024), 103340. doi:10.1016/j.ijhcs.2024.103340

[7] David Wood, Jerome S. Bruner, and Gail Ross. 1976. The Role of Tutoring in Problem Solving. *Journal of Child Psychology and Psychiatry* 17, 2 (4 1976), 89–100. doi:10.1111/j.1469-7610.1976.tb00381.x

[8] Ziqing Xu and Nick Bryan-Kinns. 2025. DeformTune: A Deformable XAI Music Prototype for Non-Musicians. doi:10.48550/ARXIV.2508.00160