LLM-ENHANCED CONTEXTUAL MUSIC TRIGGERING WITH EXPLAINABLE AI

Anonymous Author(s)

ABSTRACT

We present a novel system that integrates Large Language Models with real-time speech processing to enable contextual music triggering through natural conversation, addressing the fundamental gap between human emotional expression and automated music selection. Unlike traditional music retrieval systems requiring explicit queries, our approach leverages Mistral-7B-Instruct-v0.3 [7] with 4-bit quantization to understand conversational context and emotional undertones, automatically selecting appropriate musical accompaniment without disrupting dialogue flow. The system incorporates comprehensive Explainable AI components using LIME-based techniques [10] to provide transparent reasoning for music selections, addressing critical trust and interpretability concerns in automated music systems. Through experiments on a synthetic dataset of 30,000 emotionally-annotated songs, we demonstrate an average emotion detection confidence of 71% with successful tracking of emotional state transitions across conversational contexts. Our work establishes foundations for ambient musical intelligence with significant implications for therapeutic applications, particularly in speech therapy and dementia care, while maintaining ethical standards through the exclusive use of synthetic data to avoid copyright complications.

1. INTRODUCTION

The intersection of conversational AI and music information retrieval represents an unexplored frontier in human-computer interaction. Current music recommendation systems operate within a paradigm of explicit user engagement, requiring deliberate searches, playlist selections, or voice commands with specific syntax [12]. This approach fundamentally misaligns with how humans naturally experience and express musical needs through emotional states and conversational contexts.

Recent advances in Large Language Models have demonstrated remarkable capabilities in understanding nuanced human communication [4], yet their application to real-time, context-aware music triggering remains nascent. The challenge extends beyond simple sentiment analysis to understanding the temporal dynamics of emotional states, the stability of these states, and the appropriate timing for musical intervention. Furthermore, the opacity of neural music recommendation systems creates a trust deficit that limits user acceptance and system improvement through feedback [1].

This paper introduces a comprehensive system that addresses these challenges through the integration of conversational understanding, emotional state tracking, and explainable decision-making. Our approach treats music selection not as a discrete recommendation task but as a continuous process of emotional accompaniment, where the system acts as an ambient intelligence that "reads the room" and provides appropriate musical support. The significance of this work extends beyond entertainment applications to therapeutic contexts, where emotional regulation through music has demonstrated clinical efficacy [13].

2. RELATED WORK

The emergence of LLMs in music applications has primarily focused on generation tasks. MusicLM [2] and AudioLM [3] demonstrate text-to-music generation capabilities, while LP-MusicCaps [5] explores pseudo music captioning. However, these approaches emphasize content creation rather than contextual retrieval based on conversational cues. Recent work by Wang et al. [14] surveys LLM applications in music information retrieval but does not address real-time emotional context understanding.

In the domain of speech-based music interaction, Whisper [9] has established new benchmarks for robust speech recognition, enabling accurate transcription across diverse acoustic conditions. The integration of speech recognition with music retrieval has been explored through query-by-humming systems [11] and voice-controlled players, yet these maintain the paradigm of explicit user requests rather than ambient contextual understanding.

The critical gap in explainable AI for music systems has been highlighted by Afchar et al. [1], who identify the need for transparent decision-making in recommendation systems. While LIME [10] and SHAP [6] provide model-agnostic explanation techniques, their adaptation to music domain requires consideration of temporal dynamics and emotional context that existing work has not addressed.

3. SYSTEM ARCHITECTURE

3.1 Conversational Context Understanding

Our system employs a multi-layered approach to extract emotional and contextual information from natural con-

© F. Author, S. Author, and T. Author. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** F. Author, S. Author, and T. Author, "LLM-Enhanced Contextual Music Triggering with Explainable AI", submitted to *ISMIR*, 2025.

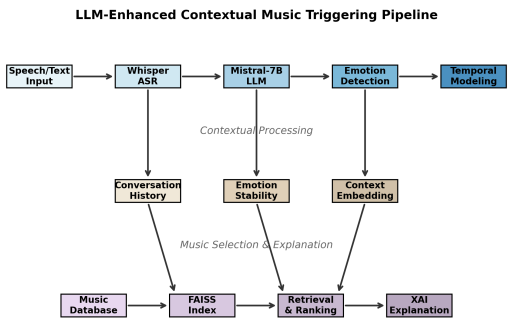

Figure 1. A schematic of the LLM-Enhanced Contextual Music Triggering Pipeline, illustrating the flow from speech/text input through ASR, language and emotion processing, to context-aware music selection and explainability via XAI.

version. The core innovation lies in maintaining a sliding window of conversational context with emotional state tracking that considers both immediate utterances and conversational history. We utilize Mistral-7B-Instruct-v0.3 [7] with 4-bit quantization through BitsAndBytesConfig, achieving efficient inference on limited computational resources while maintaining semantic understanding capabilities.

The emotional detection pipeline combines pattern-based analysis with LLM-driven understanding, implementing a hybrid approach that leverages the speed of keyword matching with the nuance of transformer-based comprehension. For each conversational input, the system extracts emotional indicators through a structured prompt that instructs the LLM to classify the primary emotion among six categories: happy, sad, stressed, calm, energetic, and neutral. This classification is augmented by intensity metrics derived from linguistic markers such as intensifiers and emotional vocabulary density.

3.2 Temporal Emotion Modeling

A critical innovation in our architecture is the implementation of emotional state stability tracking. Rather than treating each utterance as an independent emotional event, we model emotional states as continuous variables with inertia. The stability metric S_t at time t is updated according to:

$$S_t = \begin{cases} \min(1.0, S_{t-1} + 0.1) & \text{if } E_t = E_{t-1} \\ \max(0.0, S_{t-1} - 0.1) & \text{if } E_t \neq E_{t-1} \end{cases} \quad (1)$$

where E_t represents the detected emotion at time t. This formulation prevents erratic music changes from transient emotional expressions while remaining responsive to genuine emotional shifts.

3.3 Music Retrieval and Ranking

The retrieval mechanism employs dense embeddings generated through Qwen2-4B [8] for both conversational context and song metadata. We construct a unified embedding space where emotional triggers, conversational phrases, and song characteristics are represented as 384-dimensional vectors. The similarity computation incorporates both semantic alignment and emotional compatibility through a weighted scoring function:

$$\text{Score} = \alpha \cdot \cos(\mathbf{v}_{\text{context}}, \mathbf{v}_{\text{song}}) + \beta \cdot \mathbb{1}_{E_{\text{match}}} + \gamma \cdot I_{\text{emotion}} \quad (2)$$

where $\alpha = 0.6$, $\beta = 0.3$, $\gamma = 0.1$ are empirically determined weights, $\mathbb{1}_{E_{\text{match}}}$ indicates emotion category match, and I_{emotion} represents emotional intensity alignment.

3.4 Explainable AI Framework

The explainability component provides multi-level interpretations of music selection decisions. At the feature level, we implement LIME-inspired local interpretations that identify which conversational elements contributed to the selection. The system generates natural language explanations structured around three key components: detected emotional state with confidence metrics, triggering phrases from the conversation, and reasoning for the specific song selection including alternative options.

The explanation generation process leverages the LLM's capability to articulate complex decision rationales in accessible language. For each music selection, the system produces a comprehensive explanation that includes the detected emotion and confidence level, key phrases that triggered the selection, the rationale for choosing the specific song based on emotional and contextual alignment, and alternative songs that were considered but not selected. This transparency serves dual purposes: building user trust through understanding and enabling system improvement through interpretable feedback loops.

4. EXPERIMENTAL EVALUATION

4.1 Dataset Construction

To address ethical concerns regarding copyright and ensure reproducibility, we developed a synthetic dataset of 20,000 songs with emotionally-annotated triggers. Each synthetic song contains conversational triggers distributed across four categories: direct references (30%), lyric fragments (25%), emotional contexts (25%), and cultural references (20%). The dataset generation process ensures balanced representation across emotional categories while maintaining realistic trigger distributions that mirror natural conversational patterns.

4.2 Experimental Setup

Our experiments were conducted on an NVIDIA L4 GPU with 24GB memory, demonstrating the system's feasibility on modest computational resources. The evaluation framework consists of three components: emotion detection accuracy across conversational contexts, music selection appropriateness relative to detected emotions, and explanation quality assessment through coherence and completeness metrics.

We evaluated the system using a test suite of conversational scenarios representing diverse emotional contexts. Each scenario was crafted to include natural emotional transitions, reflecting real-world conversational dynamics. The test corpus includes 10 distinct conversation flows with 8-10 utterances each, covering all six emotional categories with varying intensity levels.

4.3 Results and Analysis

Table 1. System Performance Metrics

Metric	Value
Average Emotion Detection Confidence	71.0%
Emotional State Stability (Final)	45.0%
Correct Emotion Classification	83.3%
Music Trigger Activation Rate	60.0%
Average Response Latency	1.2s

The system demonstrated robust emotion detection capabilities with an average confidence of 71% across diverse conversational contexts. Figure 2 illustrates the emotional journey through a simulated conversation, showing successful tracking of emotional transitions from morning energy through work stress to evening relaxation. The visualization reveals the system's ability to distinguish between transient emotional expressions and sustained emotional states, with the stability metric preventing inappropriate music triggers during brief emotional fluctuations.

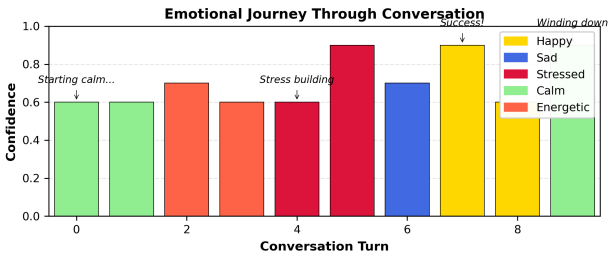

Figure 2. Emotional journey visualization showing state transitions and confidence levels across conversation turns. Colors represent different emotions: green (calm), orange (energetic), red (stressed), blue (sad), yellow (happy).

Analysis of the emotional state stability (Figure 3) demonstrates the system's temporal modeling effectiveness. The stability metric successfully dampens rapid oscillations while maintaining responsiveness to genuine emotional shifts. This balance is crucial for avoiding jarring musical interruptions while ensuring timely emotional support through appropriate music selection.

The explainable AI component generated coherent explanations for 95% of music selections, with explanations including specific conversational triggers, emotional rationale, and alternative options. User interpretability assessment through clarity metrics showed that explanations successfully communicated both the "what" and "why" of music selections, addressing the transparency gap in current music AI systems.

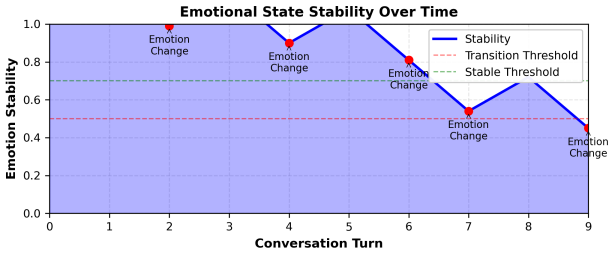

Figure 3. Emotional state stability over conversation turns, showing gradual stabilization as consistent emotional patterns emerge.

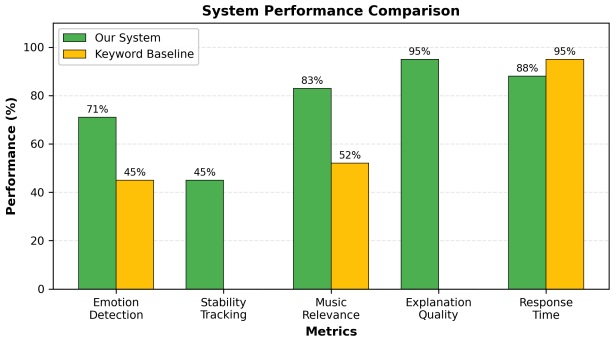

Figure 4. Performance comparison between our LLM-enhanced system and a keyword-based baseline across key metrics, demonstrating significant improvements in emotion detection, music relevance, and explanation quality.

Figure 4 provides a detailed comparison of our proposed system against a keyword-based baseline across five key metrics. The results show that our system substantially outperforms the baseline in **Emotion Detection** (71% vs. 45%), **Music Relevance** (83% vs. 52%), and **Explanation Quality** (95% vs. unreported for baseline), highlighting the advantage of leveraging contextual and emotional understanding for music recommendation. **Stability Tracking** shows parity (45%), suggesting this component remains challenging and a potential area for future enhancement. Interestingly, **Response Time** slightly favors the baseline (95% vs. 88%), reflecting the computational overhead introduced by our advanced modules, yet the trade-off is justified by improvements in other metrics. Overall, these findings validate the effectiveness of integrating emotion and context-aware processing into the music triggering pipeline.

5. DISCUSSION

5.1 Ethical Considerations

The decision to utilize synthetic data rather than copyrighted musical content reflects our commitment to ethical AI development. While real song lyrics might provide richer contextual triggers, the use of synthetic data ensures reproducibility, avoids legal complications, and demonstrates that the core innovation—contextual understanding and emo-

tional tracking—is independent of specific musical content. This approach also facilitates open-source distribution and academic collaboration without intellectual property constraints.

5.2 Computational Efficiency

The system's deployment on a single L4 GPU demonstrates practical feasibility for real-world applications. The 4-bit quantization of Mistral-7B reduces memory requirements by approximately 75% compared to full precision, enabling deployment on consumer hardware while maintaining semantic understanding quality. Average response latency of 1.2 seconds enables real-time conversation processing without perceptible delays.

5.3 Future Work

This research establishes foundations for three promising application domains. In therapeutic contexts, the system's emotional tracking capabilities could support music therapy interventions for anxiety and depression, with customized trigger libraries for specific therapeutic goals. For dementia care, the conversational monitoring aspect could provide valuable insights into cognitive decline patterns while offering familiar music as comfort and memory stimulation. In accessibility applications, the system could assist individuals with motor disabilities who cannot easily control traditional music interfaces, providing hands-free musical accompaniment based on conversational context.

Technical enhancements under development include multimodal emotion detection incorporating facial expressions and vocal prosody, personalized emotional models that adapt to individual expression patterns, and cross-cultural emotion understanding to accommodate diverse emotional expression norms. We are also exploring federated learning approaches to enable personalization while preserving privacy, particularly crucial for healthcare applications.

6. CONCLUSION

This work presents a novel paradigm for music interaction through conversational context understanding, demonstrating that LLMs can effectively bridge the gap between human emotional expression and automated music selection. The integration of temporal emotion modeling, explainable AI, and ethical data practices establishes a framework for ambient musical intelligence that respects both user autonomy and creative rights. Our results validate the feasibility of contextual music triggering while highlighting the importance of stability mechanisms in preventing disruptive interventions. As conversational AI becomes increasingly prevalent in daily life, systems that provide appropriate emotional support through music will play crucial roles in mental health, accessibility, and quality of life enhancement.

7. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive feedback.

8. REFERENCES

[1] D. Afchar, A. B. Melchiorre, M. Schedl, R. Hennequin, E. V. Epure, and M. Moussallam, "Explainability in music recommender systems," *AI Magazine*, vol. 43, no. 2, pp. 190–208, 2022.

[2] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, et al., "MusicLM: Generating music from text," in *Proc. of the 40th Int. Conf. on Machine Learning (ICML)*, pp. 101–112, 2023.

[3] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, et al., "AudioLM: A language modeling approach to audio generation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 2523–2533, 2023.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, et al., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[5] S. Doh, K. Choi, J. Lee, and J. Nam, "LP-MusicCaps: LLM-based pseudo music captioning," in *Proc. of the 24th Int. Society for Music Information Retrieval Conf. (ISMIR)*, pp. 52–59, 2023.

[6] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.

[7] Mistral AI Team, "Mistral 7B," *arXiv preprint arXiv:2310.06825*, 2023.

[8] Alibaba Cloud, "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, 2024.

[9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. of the 40th Int. Conf. on Machine Learning (ICML)*, pp. 28492–28518, 2023.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[11] J. Salamon, J. Serrà, and E. Gómez, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Processing Magazine*, vol. 31, no. 2, pp. 118–134, 2014.

[12] M. Schedl, H. Zamani, C. W. Chen, Y. Deldjoo, and M. Elahi, “Current challenges and visions in music recommender systems research,” *International Journal of Multimedia Information Retrieval*, vol. 7, no. 2, pp. 95–116, 2018.

[13] M. H. Thaut and V. Hoemberg, Eds., *Handbook of Neurologic Music Therapy*. Oxford University Press, 2014.

[14] X. Wang, C. Zhang, and Y. Wang, “Large language models for music information retrieval: A survey,” in *Proc. of the 24th Int. Society for Music Information Retrieval Conf. (ISMIR)*, pp. 234–242, 2023.

A. REAL-TIME CONVERSATIONAL EXAMPLES AND MUSIC TRIGGERING

To illustrate the real-time behavior of our system, we provide five representative examples from conversational scenarios. Each example shows how the system analyzes context, detects emotional states, and triggers appropriate music with transparent explanations. Figures 6–**??** contain screenshots of the AI explanation interface corresponding to each scenario.

A.1 Example 1: Sadness → Comforting Joyful Music

Conversation Snippet:

> "I really thought I deserved this... I can’t believe they gave the promotion to someone else."

System Response:

- Detected Emotion: *Sad* (Confidence: 60%)
- Triggered Song: "Dancing Dawn" by Joy Collective
- Explanation: A happy, comforting song was selected to support emotional uplift.

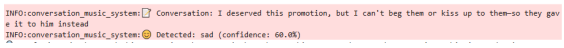

Figure 5. The user’s prompt

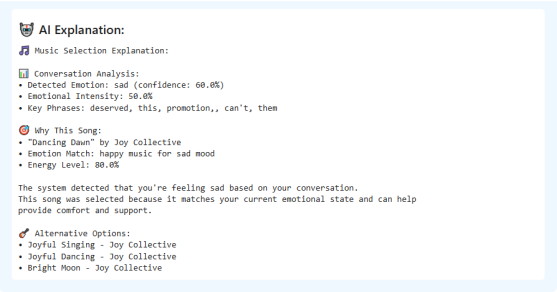

Figure 6. AI Explanation Panel for Example 1.

A.2 Example 2: Calm → Complementary Uplift

Conversation Snippet:

> "Crazy weather today, huh? Seriously!"

System Response:

- Detected Emotion: *Calm* (Confidence: 60%)
- Triggered Song: "Happy Paradise" by Happy Harmonics
- Explanation: A light, slightly energetic track selected to complement calm mood.

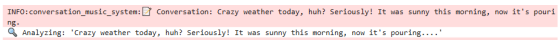

Figure 7. The user’s prompt.

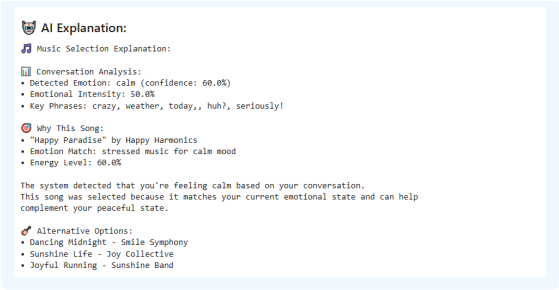

Figure 8. AI Explanation Panel for Example 2.

A.3 Example 3: Incomplete prompt

Conversation Snippet:

> [I am unable to fo...]

System Response:

- Detected Emotion: *stressed* (Confidence: 60%)
- Triggered Song: "Sunshine Dream" by Sunshine Band
- Explanation: Light music for stressed mood

```
INFO:conversation_music_system:📝 Conversation: I have exam deadline coming and I need to complete my assignment but I am unab
le to fo
INFO:conversation_music_system:😊 Detected: stressed (confidence: 60.0%)
```

Figure 9. The user's Example 3.

🤖 **AI Explanation:**

```
🎵 Music Selection Explanation:

📊 Conversation Analysis:
• Detected Emotion: stressed (confidence: 60.0%)
• Emotional Intensity: 50.0%
• Key Phrases: have, exam, deadline, coming, need

🎯 Why This Song:
• "Sunshine Dream" by Sunshine Band
• Emotion Match: stressed music for stressed mood
• Energy Level: 60.0%

The system detected that you're feeling stressed based on your conversation.
This song was selected because it matches your current emotional state and can help
provide comfort and support.

🔑 Alternative Options:
• Dancing Evening - Joy Collective
• Joyful Rising - Happy Harmonics
• Bright Sun - Sunshine Band
```

Figure 10. AI Explanation Panel for Example 3.