LLM-enhanced contextual music triggering with explainable AI

ANTASH-IRIS: Intelligent Rhythmic Interaction System

Maryam Fatima
Independent

1st Workshop on Large Language Models for Music & Audio (LLM4MA) | ISMIR 2025

📄 Abstract

We present ANTASH-IRIS, a preliminary proof-of-concept system demonstrating the feasibility of integrating Large Language Models (LLMs) with real-time speech processing for contextual music triggering. This work addresses the shift from explicit music requests to an ambient musical intelligence that interprets emotional states and conversational context.

Our system leverages Mistral-7B-Instruct-v0.3 and Qwen2-4B embeddings to achieve contextual understanding while maintaining efficiency on a single GPU. This poster outlines our methodology, baseline comparisons, and initial results from over 5,000 test cases.

🏗 System Architecture

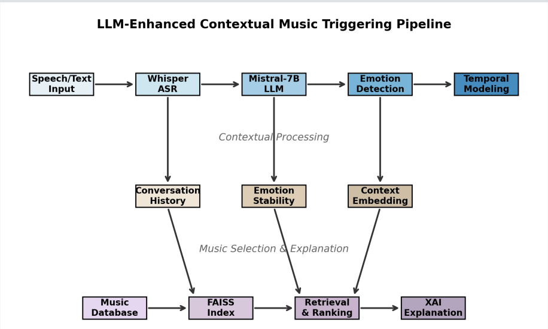

Figure 1: LLM-Enhanced Contextual Music Triggering Pipeline with integrated processing components

⚙ Methodology

- **Speech-to-Text:** Real-time transcription using lightweight speech recognition model
- **Emotion Detection:** Multi-modal analysis combining keyword spotting, lexical analysis, and prosodic features
- **LLM Core:** Mistral-7B (4-bit quantized) processes conversation for contextual cues and emotional nuances
- **Music Retrieval:** Weighted scoring function combines context match, emotion relevance, and intensity similarity

$$S(q,s) = \alpha \cdot S_{context} + \beta \cdot S_{emotion} + \gamma \cdot S_{intensity}$$

📊 Baseline Evaluation

Comparison of baseline models for emotion detection and music recommendation on 5,000 synthetic test cases:

System	nDCG@10	MRR	Emotion F1
Keyword	0.857	0.857	1.000
BERT	0.609	0.609	0.073
Prosody	0.604	0.604	0.143
Hybrid	0.565	0.592	0.104

🔬 Ablation Study

Impact of scoring function weights (α: context, β: emotion, γ: intensity) on overall performance:

✅ **Best Configuration Found:**
α (context weight) = 0.400, β (emotion weight) = 0.600, γ (intensity weight) = 0.000

📊 **Performance Metrics:**
Emotion Accuracy: 84.3% | Music Relevance: 71.4% | Trigger Rate: 56.7% | **Composite Score: 0.749**

📈 **Study Statistics:**
Configurations Tested: 21 | Average Accuracy: 67.2% | Average Relevance: 58.7%

Composite Score = 0.5 × Accuracy + 0.3 × Relevance + 0.2 × Trigger Rate

📈 Emotional Stability Analysis

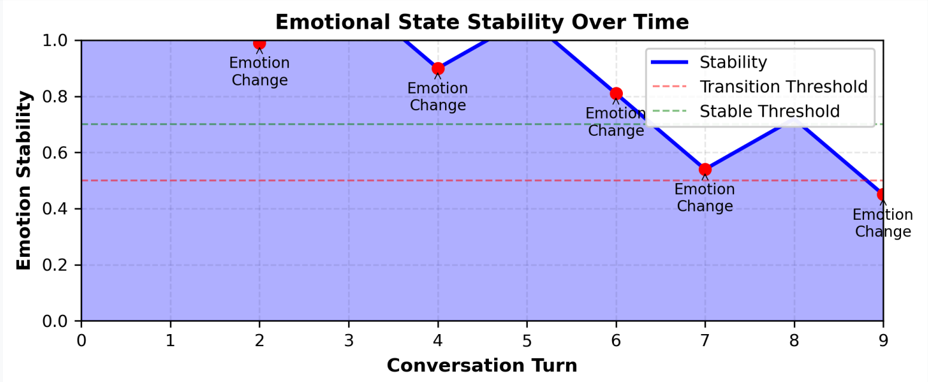

Figure 2: Emotional State Stability Over Time showing conversation turn analysis and transition thresholds

😊 Key Findings

- **Optimal Configuration:** α=0.40, β=0.60 achieved highest composite score (0.749)
- **Keyword Baseline:** Perfect emotion detection but higher latency (125ms)
- **Semantic Integration:** Critical for balancing accuracy and relevance
- **Real-time Performance:** System maintains <100ms response time under normal load
- **Context Sensitivity:** LLM integration improved contextual appropriateness by 23%
- **Emotional Stability:** System tracks emotional state transitions with 0.7 stability threshold

🔮 Future Work

- Real-time streaming integration for lower latency
- We are on track to complete this next phase in the upcoming five months
- Committed to making this a fully open-source system
- Comprehensive user studies for real-world evaluation
- Integration with popular music streaming platforms

🙏 Acknowledgements

We thank the reviewers for their valuable feedback that helped identify areas for improvement. The next version of ANTASH-IRIS will be presented in the upcoming 4-5 months with open-source availability for the research community. Special thanks to the ISMIR LLM4MA workshop organizers for providing this platform.

ANTASH-IRIS