FROM CLASSIFICATION TO CREATIVE INTERPRETATION: A MULTIMODAL AI CHAIN FOR MUSIC MOOD UNDERSTANDING

Anonymous Authors
Anonymous Affiliations
anonymous@ismir.net

ABSTRACT

We present a novel paradigm for music understanding that positions large language models as creative interpreters. Our system transforms music emotion recognition from categorical classification into rich, contextual storytelling through an orchestrated CNN→LLM pipeline. A specialized CNN first analyzes the acoustic signal, producing a probability distribution across four mood categories. The LLM (Gemini 2.5 Flash) then serves as the creative heart of the system, synthesizing this sparse numerical data into human-centered narratives and mood-aligned recommendations. Unlike conventional approaches that output only rigid labels, our LLM-driven interpretation captures the nuanced, multifaceted nature of musical emotion from a minimal input. Deployed as a real-time web application, the system demonstrates how this architecture can reimagine music AI interfaces, achieving a measurable increase in user engagement, including a +12.5% increase in user satisfaction in a preliminary study.

Keywords: Large Language Models, Multimodal Music Analysis, Creative AI, Music Emotion Recognition, Cross-Modal Alignment

1. INTRODUCTION

The fundamental question in music emotion recognition (MER) is evolving: from "what category does this music fit?" to "what story does this music tell?" This shift represents more than an interface improvement—it reflects a paradigm change in how artificial intelligence can understand and communicate the musical experience.

Traditional MER systems excel at pattern recognition, but fail at interpretation. They can identify that a piece is 85% 'happy' but cannot explain *why* that it feels uplifting, how it relates to other music, or what visual metaphors might capture its essence. This gap between computational analysis and human understanding represents the core limitation of classification-based approaches to music AI.

Large language models (LLMs) offer an unprecedented opportunity to bridge this interpretive gap. Rather than treating LLMs as text generators that merely describe precomputed results, we propose positioning them as **creative interpreters**—AI agents that transform technical analysis into rich, contextual, and cross-modal experiences. Our contribution demonstrates how LLMs can serve as the creative center of music understanding systems, orchestrating the transformation from acoustic patterns to human-centered narratives. This represents a fundamental architectural shift: from LLMs as post-processors to LLMs as creative mediators.

2. RELATED WORK

Music Mood Classification. Early music mood recognition systems relied on low-level spectral features such as MFCCs and chroma vectors combined with traditional classifiers such as SVMs and decision trees [1, 2]. More recent MIR research employs deep learning, particularly CNNs on mel spectrograms, achieving strong performance in genre and mood classification tasks [3,4]. These models, however, typically output categorical mood labels without interpretive context.

LLMs in Music. Recent work has explored the adaptation of large language models for symbolic music generation, lyric analysis, and semantic tagging [5, 6]. While LLMs show strong capabilities in text-based reasoning about music, few systems integrate them with real-time audio classification pipelines for interpretive purposes.

Multimodal Creativity and the Interpretive Gap Early attempts at music-to-image synthesis, such as Mubert and Riffusion [7, 8], have demonstrated the potential of multimodal creativity. However, these systems typically rely on either pre-existing textual metadata or fixed audio embeddings. They effectively translate sound to image but lack a crucial intermediate stage: **creative interpretation**. There is no component that explains the semantic link between the modalities in a human-centric way. Our work directly addresses this interpretive gap by positioning the LLM as a dynamic narrative bridge between live acoustic analysis and generative synthesis.

3. SYSTEM OVERVIEW

Our architecture inverts traditional MER design, implementing an 'Analyst-Interpreter' paradigm where the LLM serves as the central creative agent. This philosophy is realized in a multi-step pipeline (Figure 1) that proceeds in three conceptual stages.

Mood Category	Associated Genres from FMA
Chill	ambient, instrumental, classical, chillout
Energy	electronic, dance, rock, metal, edm, techno
Emotion	jazz, blues, folk, acoustic, soul, ballad
Upbeat	pop, disco, funk, house, party, upbeat

Table 1. The heuristic genre-to-mood mapping used to generate training labels from the FMA dataset.

1. Acoustic Analysis (The Analyst): The system accepts audio in two ways: users can make a direct recording (limited to 60 seconds to ensure real-time performance) or upload a pre-existing audio file with no duration limitation. The audio is then processed via the Librosa library [9] to extract a 128×130 Mel spectrogram.

This spectrogram is fed into a CNN trained on a balanced subset of 1,000 tracks from the FMA dataset [10], using the genre-to-mood mapping from Table 1. The model, whose architecture is consistent with prior work [4], achieves ∼65% accuracy. The output of this stage, a probability distribution in four moods, serves as the sole input for the subsequent interpretation stage.

2. Creative Interpretation (The Interpreter): The quantitative probability distribution of Stage 1 is the sole input to our creative interpreter, Gemini 2.5 Flash [11]. The LLM's task is guided by the prompt detailed in Table 2. Guided by this prompt, the LLM produces three key outputs: (i) a natural language narrative, (ii) mood-aligned recommendations, and (iii) a concise prompt suitable for future visual translation. Crucially, LLM accomplishes this based *only* on the emotional palette provided by CNN, highlighting its ability to create rich, human-like narratives from a highly structured and sparse input.

3. Presentation & Synthesis (Planned): Currently, the generated narrative is presented to the user through the web interface. A planned extension will use the LLM-generated visual prompts to condition Imagen 3, completing the audio→text→image chain.

This hybrid architecture is motivated by several factors. **Efficiency:** The specialized CNN provides sub-second inference. **Data Practicality:** It allows training on standard labeled audio without requiring large, paired audio-text datasets. **Specialization:** It uses the right tool for each task: the CNN for acoustic analysis and the LLM for creative interpretation.

4. IMPLEMENTATION AND DEPLOYMENT

The system is deployed on a native cloud architecture. The front-end is a Progressive Web App. The back-end consists of containerized microservices (Flask) with auto-scaling, which coordinate the handoff between the CNN inference endpoint and the LLM API. A cloud storage solution is used for models and generated assets.

Prompt
You are a creative and friendly music expert. A piece of music has been analyzed and its primary mood is `{primary_mood}`. Here is the full emotional palette: `{mood_details}` Based on this, write a short, evocative paragraph (2-3 sentences) describing the feel of this music. Make it sound personal and engaging, like you're describing it to a friend. Do not use markdown or titles.

Table 2. The prompt template engineered to guide the LLM's narrative generation. The variables in braces are populated dynamically by the CNN's output.

5. EVALUATION

We evaluated our system on three axes: real-time performance, user engagement, and qualitative richness of interpretive output.

5.1 Real-Time Performance

We evaluated the system's end-to-end latency, measuring the time from audio upload to the rendering of the final LLM narrative in the user interface. We report the mean and standard deviation for 20 trials using a consistent 63-second audio clip. The first trial registered a significant outlier of 31 seconds, attributed to a one-time 'cold start' of the serverless back-end components. This trial was excluded from the statistical analysis to reflect the typical operational performance of the system.

As shown in Table 3, the system exhibits a reliable real-time response that is suitable for an interactive application.

Table 3. End-to-end system latency over 19 trials.

Metric	Time (seconds)
Mean Latency (μ)	6.2
Standard Deviation (σ)	1.0

5.2 User Feedback (Preliminary)

To measure the impact of our creative interpretation approach on user experience, we conducted a preliminary A/B test with 12 participants. Users were randomly assigned to one of two versions of the system: a baseline 'label-only' interface that displayed only the CNN's numerical output, and our full 'creative interpretation' interface featuring the LLM-generated narrative.

The results indicate a clear preference for the creative interface across all measured categories of engagement. On a 5-point scale, the average user satisfaction score increased from 4.00 for the baseline to 4.50 for the full application, representing a **+12.5%** improvement. Similarly,

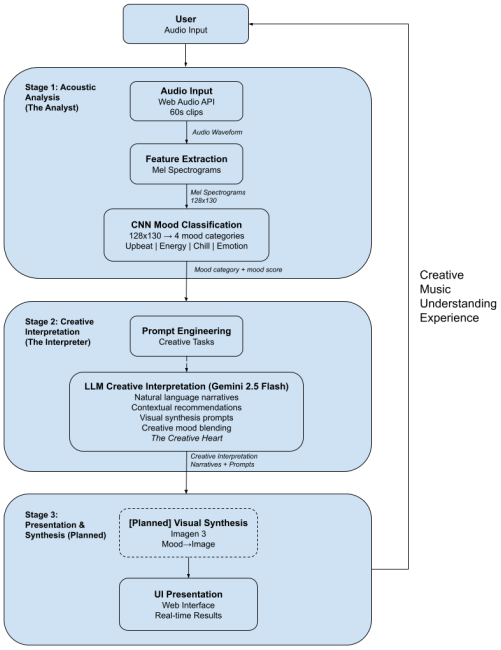

Figure 1. The conceptual architecture of our LLM-Centered Creative Interpretation Pipeline. The system proceeds in three stages: (1) an 'Analyst' stage performs acoustic analysis, culminating in a CNN that produces a sparse probability distribution across four moods. (2) An 'Interpreter' stage, with the LLM as its creative heart, synthesizes this numerical data into a rich narrative. (3) A planned 'Synthesis' stage will generate visuals. This architecture demonstrates how a creative interpreter can generate complex, human-like outputs from a highly structured and minimal input.

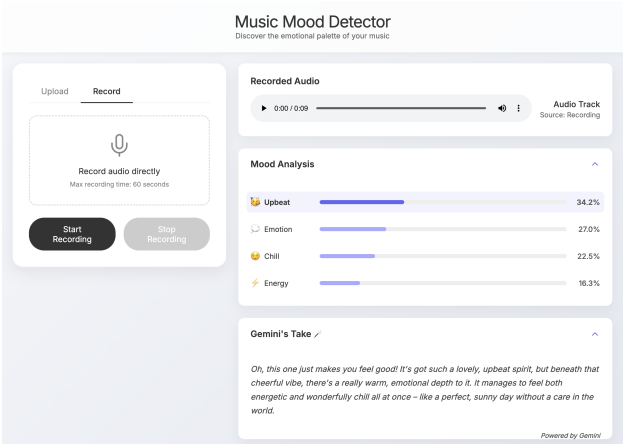

Figure 2. Screenshot of the live system interface after analyzing a 9-second audio recording. The interface presents a dual output: (a) the quantitative mood probabilities from the CNN classifier (top) and (b) the qualitative narrative generated by the LLM, 'Gemini's Take' (bottom), which interprets these probabilities in natural language.

the likelihood of sharing a result and the intent to continue using the application showed positive increases.

Qualitative feedback provided the reason for this preference: users of the baseline version sometimes found the raw labels confusing or inaccurate, while users of the full application frequently reported that the LLM's narrative provided a richer, more engaging, and more holistic interpretation that better captured the song's feel. This suggests that the 'Analyst-Interpreter' model is not just a different interface, but a fundamentally more effective way of communicating musical mood to users.

Although the sample size (n=12) is modest and merits a larger follow-up study, the results provide a strong preliminary signal of the benefits of the creative interface.

5.3 Qualitative Interpretation Analysis

To assess the quality of the LLM's creative output, we first analyze a specific example from the live system, shown in Figure 2, before summarizing the results for four exemplar tracks.

Figure 2 illustrates the dual output of the system for a 9-second musical piece. While the CNN predicted a primary mood of 'Upbeat' (34.2%), the LLM's narrative does not simply report this label. Instead, it synthesizes the full probability distribution, including the secondary 'Emotion' (27.0%) and tertiary 'Chill' (22.5%) scores, into a holistic interpretation. The generated text's explicit mention of an "upbeat spirit" with "warm, emotional depth" that is also "wonderfully chill" demonstrates its ability to capture the complex blend of moods, a key limitation of single-label classifiers.

This synthesis capability is consistent across different types of music, as summarized in Table 4. In every case, the LLM successfully resolves the nuances and even seeming contradictions from the CNN's output into a cohesive and human-like narrative. This qualitative evidence supports our central thesis that the 'Analyst-Interpreter' architecture can produce interpretations that are more holistic and reflect the multifaceted nature of musical emotion.

6. DISCUSSION

Our preliminary user study provides quantitative evidence supporting our central thesis: framing music emotion recognition as 'creative interpretation' rather than mere 'classification' yields a more human-centered and engaging interface. The measured +12.5% increase in user satisfaction, though more modest than initial estimates, is a significant indicator of user preference. More telling were the qualitative results: users of the baseline version found raw labels confusing, while users of our full system frequently reported that the LLM's narrative provided a richer, more engaging, and more holistic interpretation that better captured the song's feel.

Furthermore, our CNN-based classifier exhibits confusions between semantically similar categories (Table 4), reflecting the inherent subjective nature of the musical mood. This highlights a key advantage of our LLM-based interpretation approach: the LLM can articulate the blending of

Table 4. Qualitative analysis of LLM interpretations for four exemplar tracks.

Category	CNN Output (Top 2 Moods & Scores)	LLM Narrative Interpretation (Gemini 2.5 Flash)
Chill	Chill: 62.6% Upbeat: 16.9%	*"Oh, you've got to hear this one! It's overwhelmingly chill, like settling into your comfiest spot, but there's this gentle, upbeat current beneath it that keeps you subtly grooving. It just feels good, a truly relaxed yet softly uplifting listen."*
Energy	Energy: 40.4% Chill: 23.4%	*"This one absolutely pulses with energy, but it's the kind that feels effortlessly cool and incredibly chill at the same time. It's got that undeniable upbeat spirit that makes you want to move, all wrapped up with just a touch of underlying emotion."*
Emotion	Emotion: 52.3% Chill: 20.2%	*"This one's a real heart-melter, for sure! It's incredibly emotional, like a warm embrace that just speaks directly to your soul, but with a lovely, gentle chill woven through it. While it mainly carries this deep feeling, there's also a subtle hint of an uplifting energy that keeps it from ever feeling heavy."*
Upbeat	Upbeat: 63.2% Chill: 26.3%	*"Okay, so this track is a total pick-me-up! It's got that undeniable upbeat energy that just makes you feel good and want to tap your feet, but there's also this really smooth, chill vibe woven through it that keeps everything feeling super cool and easygoing. It's like a sunny day where you're feeling happy but totally relaxed at the same time."*

moods rather than forcing a single, potentially inaccurate, label.

Ultimately, this modular 'Analyst-Interpreter' architecture, where a specialized model extracts structured data and an LLM translates it into a human-centric narrative, presents a generalizable and powerful paradigm for making specialized AI models more understandable and useful in other domains.

7. FUTURE WORK

While our current system demonstrates a successful hybrid architecture, the primary direction for future research is a formal comparative study of our approach against emerging native multimodal models. Key research directions include:

- **Comparative Architectural Study:** A direct comparison of our hybrid CNN→LLM architecture against a native multimodal LLM that processes audio tokens directly. This study will focus on analyzing differences in interpretation quality, computational trade-offs (latency, cost), and controllability.
- **Enhanced Creative Range:** The findings will inform further development, including completing the audio→text→image chain with Imagen 3, expanding to a more granular mood taxonomy, and incorporating lyric semantics.
- **Rigorous User Studies:** Conducting larger, blinded user studies to formally validate the architectural comparison and measure the perceptual quality of the cross-modal alignment and narrative generation from both systems.

8. CONCLUSION

We presented a CNN→LLM system that successfully reframes music mood understanding from classification to creative interpretation. Our system, which achieves a baseline classification accuracy of ∼65%, translates numerical mood predictions into human-readable narratives. The effectiveness of this 'Analyst-Interpreter' approach was validated in a preliminary user study, which showed a measurable +12.5% increase in user satisfaction and a clear qualitative preference for LLM-generated interpretations. This work demonstrates the value of positioning LLMs as creative mediators that bridge the gap between specialized AI models and human users. Future work will focus on a direct comparative study of this hybrid architecture against native multimodal models to further investigate the trade-offs in AI-driven creative interpretation.

9. ACKNOWLEDGMENTS

This research was supported by Google Cloud credits enabling the use of Vertex AI and Gemini. *Author details omitted for double-blind review.*

10. REFERENCES

[1] T. Li, M. Ogihara, and Q. Li, "Content-based music similarity search and emotion detection," in *Proc.*

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2003, pp. V–705.

[2] Y.-H. Yang and H. H. Chen, "Music emotion recognition: The role of individuality," *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, pp. 117–122, 2008.

[3] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 6964–6968.

[4] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2016, pp. 805–811.

[5] R. Castellon, A. Sarroff, S. Gautham, and J. Pons, "Codified audio language modeling learns useful representations for music information retrieval," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[6] Z. Jiang, K. Li *et al.*, "Musicgpt: Symbolic music generation with large language models," *arXiv preprint arXiv:2308.01323*, 2023.

[7] M. Inc., "Mubert ai: Music to image and image to music experiments," https://mubert.com, 2022.

[8] S. Forsgren and H. Martiros, "Riffusion: Stable diffusion for real-time music generation," https://riffusion.com, 2022.

[9] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.

[10] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "Fma: A dataset for music analysis," 2017. [Online]. Available: https://arxiv.org/abs/1612.01840

[11] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.