From Classification to Creative Interpretation: A Multimodal AI Chain for Music Mood Understanding

Khongorzul Munkhbat[1] Bilguun Jargalsaikhan[2] Keun Ho Ryu[3]

[1]Deep Tech LLC, Ulaanbaatar, Mongolia
khongorzul@deeptech.mn

[2]Database and Bioinformatics Laboratory, School of Electrical and Computer Engineering, Chungbuk National University, Cheongju, South Korea
bilguun@chungbuk.ac.kr

[3]Data Science Laboratory, Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam
khryu@tdtu.edu.vn

1. The Problem: Beyond Boring Labels

- Traditional Music Emotion Recognition (MER) provides rigid, uninspiring labels (e.g., 'happy').
- It fails to explain *why* a song feels uplifting or what story it tells.
- This gap limits the potential for truly creative AI music tools.

2. Our Approach: Creative Interpretation

We reframe MER from a classification task to one of **creative interpretation**. Our goal is to answer the question:

> *"What story does this music tell?"*

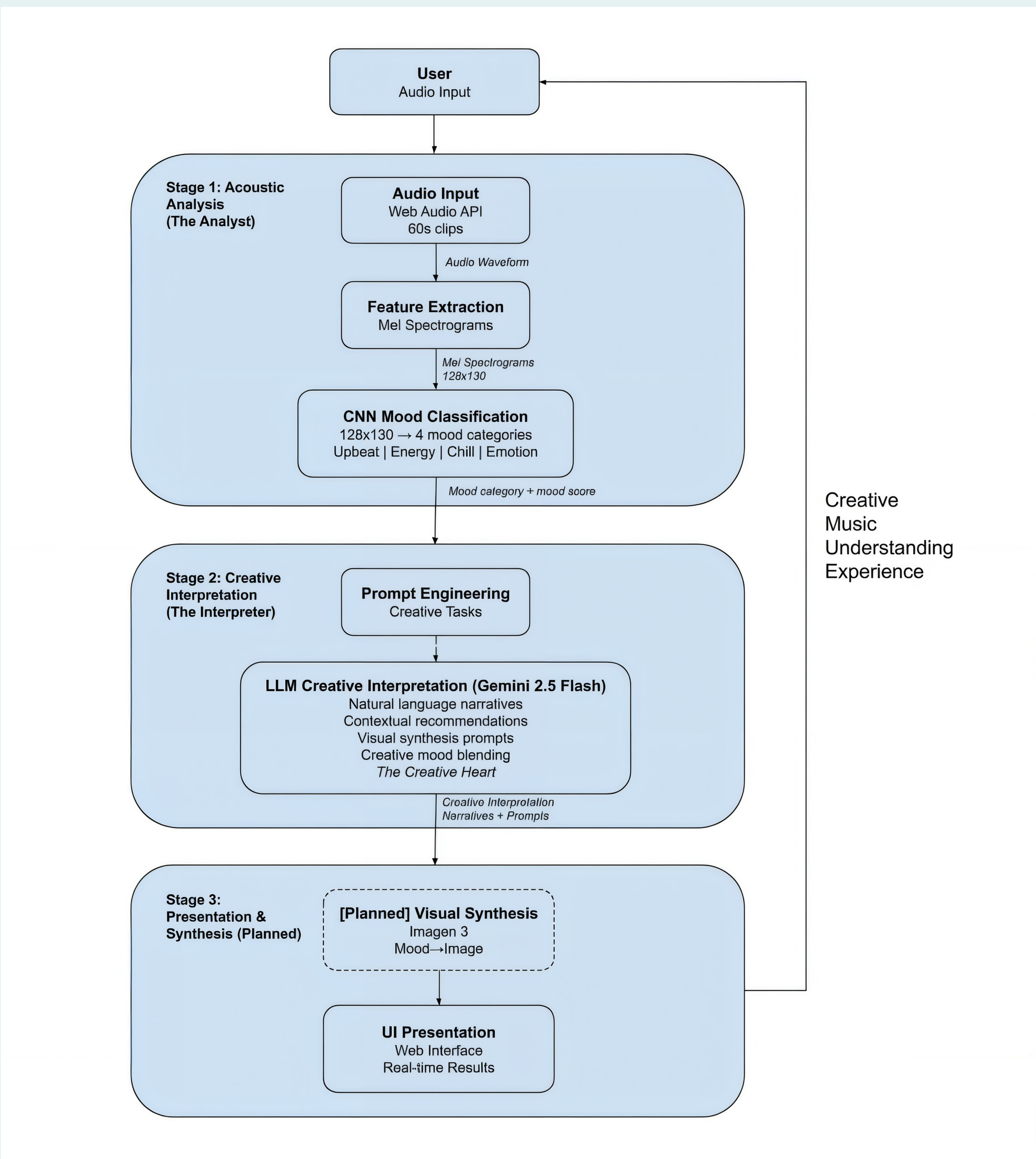
We achieve this by positioning a Large Language Model (LLM) as a **Creative Interpreter**—an AI agent that transforms technical audio analysis into rich, human-centered narratives.

3. Key Innovation: LLM as Creative Mediator

Unlike traditional post-processing approaches, our LLM serves as the **creative heart** of the system:

- Transforms sparse numerical data into rich narratives
- Captures mood complexity that single labels miss
- Bridges computation-human understanding gap

System Overview



Three-stage multimodal chain:
Analysis → Interpretation → Synthesis

4. The 'Analyst-Interpreter-Synthesizer' Pipeline

Our system is an end-to-end multimodal chain that proceeds in three stages shown in System Overview:

Stage 1: The Analyst (CNN)

A CNN analyzes the audio's Mel spectrogram to produce a nuanced **"Emotional Palette"**.

Dataset: 1,000 balanced FMA tracks across 4 mood categories:

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

Technical Implementation:

- 128×130 Mel spectrogram input via Librosa
- Genre-to-mood mapping enables probability distributions
- Achieves ~65% classification accuracy

Stage 2: The Interpreter (LLM)

Gemini 2.5 Flash synthesizes sparse numerical data into **evocative narratives** and visual prompts.

- Transforms probability distributions into human-centered stories
- Captures complex mood blends that single labels miss
- Creates mood-aligned recommendations

Example: "Chill: 62.6%, Upbeat: 16.9%" → "It's overwhelmingly chill, but there's this gentle, upbeat current beneath it that keeps you subtly grooving."

Stage 3: The Synthesizer (Planned)

Future work will use LLM-generated prompts to condition Imagen model for **mood-aligned visual art**.

5. Results: A More Engaging Experience

Real-Time Performance (19 trials):

Metric	Value
Mean Latency (μ)	6.2 seconds
Std Deviation (σ)	1.0 seconds

User Study Results (n=12, A/B Test):

+12.5% increase in user satisfaction
(Creative interpretation vs. label-only baseline)
(4.50 vs 4.00 on 5-point scale)

Key Qualitative Findings:

- Raw labels: **"confusing or inaccurate"**
- LLM narratives: **"richer, more engaging"**
- System latency: **6.2±1.0 seconds**
- Captures **complex blend of moods**
- Resolves nuances into **cohesive narratives**

6. Live System Example

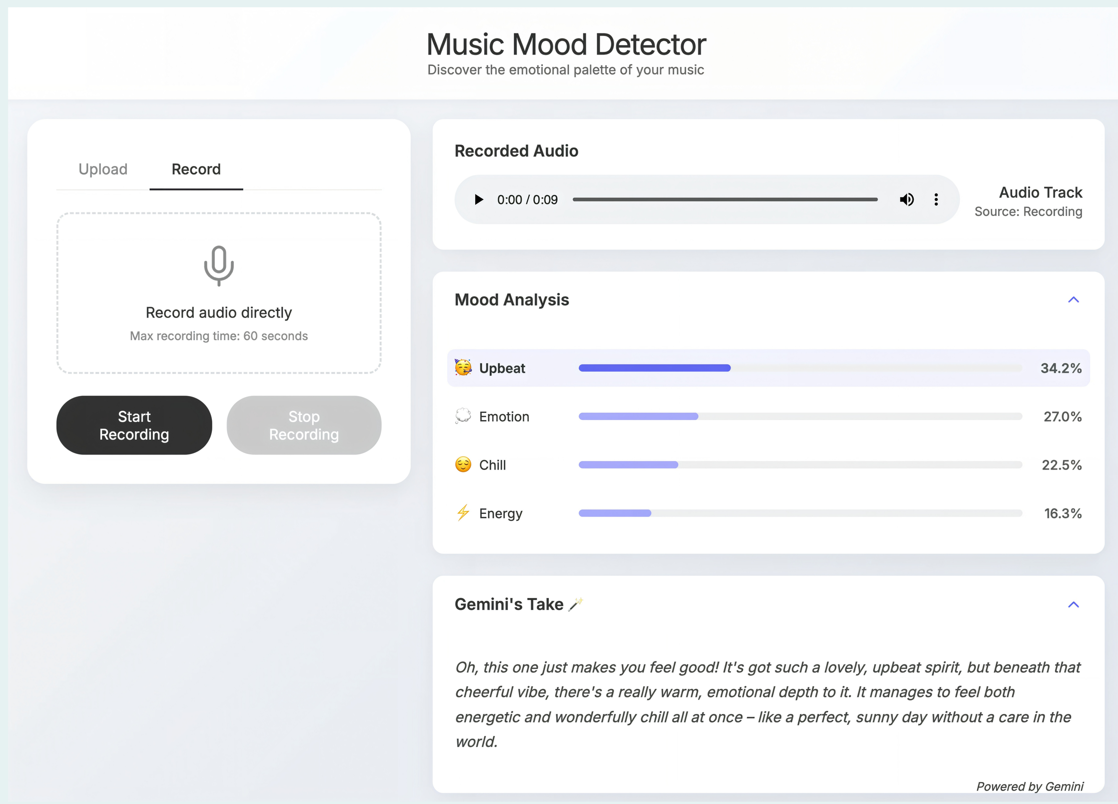


Figure: The audio→text interface output

9-second audio analysis demonstrates key innovation:
CNN Raw Output: *Upbeat (34.2%), Emotion (27.0%), Chill (22.5%), Energy (16.3%)*

LLM Creative Interpretation: *"This track has such an upbeat spirit that immediately lifts your mood, but there's also this wonderful warm, emotional depth running through it that makes it feel really meaningful. It's got this wonderfully chill vibe too that just makes the whole thing feel effortless and cool."*

Key Insight: LLM synthesizes full probability distribution into holistic interpretation, capturing musical complexity that single labels miss.

7. Qualitative Analysis: LLM Interpretation Quality

Chill Track	*"It's overwhelmingly chill, like settling into your comfiest spot, but there's this gentle, upbeat current beneath it..."*
Energy Track	*"This one absolutely pulses with energy, but it's the kind that feels effortlessly cool and incredibly chill at the same time..."*
Emotion Track	*"This one's a real heart-melter! It's incredibly emotional, like a warm embrace that speaks directly to your soul..."*
Upbeat Track	*"This track is a total pick-me-up! It's got that undeniable upbeat energy but also this really smooth, chill vibe..."*

Consistent Pattern: LLM successfully resolves seeming contradictions from CNN output into cohesive, human-like narratives.

8. Limitations & Future Work

Current Limitations:

- Small user study (n=12)
- CNN accuracy (~65%) improvable
- Visual synthesis not implemented

Future Directions:

- Complete audio→text→image pipeline with Imagen model
- Hybrid vs. native multimodal comparison
- Larger user studies (n>100)

9. Conclusion

We demonstrated an audio→text system that reframes music analysis as **creative interpretation**. The modular 'Analyst-Interpreter' architecture is a powerful paradigm for making specialized AI models more engaging and understandable. Strong positive user feedback validates this human-centric approach.

Impact & Applications: Music streaming platforms, therapeutic applications, creative tools for artists and producers.

Keywords: LLM, Multimodal Music Analysis, Creative AI, MER

Acknowledgments: This research was supported by Google Cloud credits.

Live Demo - Scan to Try!