ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

MelodyCraft: A Prompt-Based Modular AI Framework for Synchronized Lyrics and Instrumental Music Generation

Ayushi Chauhan[1], Rituraj Jain[2], Harshalkumar Vanpariya[3], Keyur Kacha[4] and Manisha Makawana[5]

ABSTRACT

Current artificial intelligence (AI) music generation systems are usually not synchronized with lyrics and melodies, emotional correlativity, or objective assessment criteria. To overcome these weaknesses, this study proposes MelodyCraft, a prompt-based modular AI architecture that combines MusicGen-small to produce instrumental music and a Mixtral transformer to create genre- and emotion-specific lyrics. The fine-tuning of QLoRA-based optimization was performed using 28,000 prompt-lyric pairs with a validation loss of 1.85 and a BLEU score of 0.65, which indicates high lyrical coherence and style fidelity. A spectral analysis of 10 human-composed and 10 AI-generated music of pop, rock, and jazz types showed no statistically significant ($p > 0.05$) spectral centroid, bandwidth, or roll-off differences among them, which is why it was deemed to have almost human acoustic realism. In addition, a human listening experiment involving 16 subjects found average Mean Opinion Scores (MOS) of 3.9-4.0 in melody quality, lyrical coherency, and emotional connection, creating perceptual parity with human compositions. MelodyCraft is a full-stack multimodal music-generating system that generates semantically, rhythmically, and emotionally consistent music and can serve as a guide for future creative systems assisted by AI.

Article information:

Keywords: Artificial Intelligence, Machine Learning, Multimodal Music Generation, Music Generation, Prompt-Based Composition

Article history:
Received: July 31, 2025
Revised: October 4, 2025
Accepted: December 25, 2025
Published: January 24, 2026
(Online)

DOI: 10.37936/ecti-cit.2026201.263332

1. INTRODUCTION

The convergence of AI and creative arts in music generation has been revolutionary in the field of computational creativity, particularly in the area of automated music generation. As AI technologies improve, systems capable of writing full musical compositions, including lyrics, melodies, and computer-generated vocals, have left behind their experimental forms and made a leap to commercial platforms used in the real world. This has been enabled by innovations in natural language processing (NLP) and deep learning, particularly transformer-based models that allow machines to learn musical forms, understand human intent, and produce coherent and emotionally expressive music.

Most current frameworks target single components, including lyrics generation, melody synthesis, and vocal rendering. This kind of fragmentation results in synchronization, emotional fit, and genre fit deficits, which create an output that is not always attuned to the desired creative essence or contextual flow. In addition, they rarely provide real-time control or user-defined parameters, such as mood, tempo, or thematic tone, restricting expressiveness and user interactivity.

This study addresses these shortcomings by proposing MelodyCraft, a prompt-based modular AI architecture that can create synchronized lyrics and instruments using natural language prompts. The main objective was to establish a unified multimodal system in which semantic, rhythmic, and emotional integration exists among the musical elements. The hypothesis of the study is as follows: when MusicGen-small, an instrumental synthesis architecture, is combined with a fine-tuned Mixtral, a lyric-generating architecture, enhanced with metadata-conscious conditioning, token-level decoding with Encodec, and

[1,2,3,4,5]The authors are with the Department of Information Technology, Marwadi University, Rajkot, Gujarat, India, E-mail: chauhanayushi107@gmail.com, jainrituraj@yahoo.com, harshalvanpariya20@gmail.com, kylestark285@gmail.com and makwanamansi1612@gmail.com

[2]Corresponding author: jainrituraj@yahoo.com

emotion-genre matching, it will be possible to create music of nearly the same quality as that produced by humans.

MelodyCraft, which methodologically consists of transformer fusion, prompt conditioning, and genre-emotion mapping, was used to achieve a state of coherence between the textual and acoustic modalities on 28,000 pairs of prompts and song lyrics. The experimental analysis proved that lyrical consistency had a BLEU score of 0.65, and there was no significant difference ($p > 0.05$) between the AI-composed and human-composed songs in terms of spectral aspects. Another study on human listening (n=16) confirmed the authenticity of the perception, and the results of that study provided a mean opinion score (MOS) of 3.9-4.0 for melody quality, lyrical coherence, and emotional resonance.

Overall, MelodyCraft created a scalable and interpretable framework for AI-assisted creative composition, a step forward in the development of multimodal music composition, providing a connection between linguistic and acoustic intelligence by introducing an integrated and emotion-adaptive pipeline.

2. LITERATURE REVIEW

Recently, methodologies for creating deep learning models, generative architectures, and evolutionary algorithms have been used to improve the creative and technical aspects of music creation. MelodyAI, created in [1], uses RNNs, LSTMs, and attention mechanisms to generate emotionally resonant melodies. On the one hand, MelodyAI brings melodic creativity forward with large datasets but not lyric synthesis and multimodal harmony. Subsequently, [2] proposed a hybrid RNN–GAN model for generating melodies that combined temporal sensitivity and adversarial generation to develop more diverse and coherent melodies; however, this was restricted to melody generation.

In the domain of melody-to-lyrics alignment, [3] proposed a prior attention mechanism together with an n-gram loss-based syllable-level transformer (EXPLING), which improved rhythmic alignment and allowed user control. In this reverse fashion of extending this paradigm, [4] proposed ConL2M, a lyrics-to-melody system incorporating Memofu, RSE, and SeqLoss for musical style control and learning of melodic structures all in one, which they finally validated using quantitative metrics and human evaluation.

The authors of [5] used mutual information with transformer architectures to improve the interpretability of music generation by enforcing a semantic alignment between the lyrics and melody. Similarly, their GAN-based system [6] utilized conditional training and Gumbel SoftMax approximations to create discrete musical features and obtain user feedback via an interactive interface.

The TBC-LSTM-GAN was introduced by [7], a TIMIT-based dataset, which stabilizes training using GumbelSoftmax to stabilize training and comprises three independent LSTM subnetworks for handling the pitch, duration, and rest periods. The benefits of this modular architecture are superior in terms of alignment and musical plausibility. Similarly, [8] explored GAN-based approaches, where LSTMGAN combined sequence modeling with adversarial training to achieve high musical coherence and considered the loss graphs as performance indicators of the model.

Rule-based systems, such as [9], are good examples of non-ML approaches that generate music from lyrics without training data. Aimed at writing singable melodies, especially for children, it is rooted in music theory and adjusts time signatures and tempos to create melodies. Researchers in [10] presented LYRA, a hierarchical framework in which text-based training was decoupled from melody-constrained inference for an unsupervised generation. Although generalization to low-resource languages is still limited, LYRA shows significant performance improvements over general lyric generation methods, such as SongMASS, because it is a rhythm-aware lyric generation approach.

These advancements are contextualized in broader overviews, including [11], [12]. The authors of [11] described the progression from RNNs to MusicVAE and explored commercial tools such as Flow Machines and Amper Music, acknowledging the lack of emotional expressiveness and the apparent ethical issues. In addition, [12] asserted that AI has the power to democratize the act of music creation and further discussed Transformers, LSTMs, and GANs as essential enablers of efficiency and accessibility.

In [13], MMGen is a melody-guided diffusion model for producing melodies based on audio and text, where melodies are aligned with audio and text by introducing a new MusicSet dataset. In terms of both style and content integration, MMGen has great strengths, although more work is needed to scale it to other genres. In contrast, [14] adopted a DIWFA GAN, which enhances emotion expression and harmonic coherence by utilizing swish activations and a GTO optimizer. On the MIDI dataset, it significantly outperformed the baselines on multiple metrics.

The author in [15] described how the use of AI procedural composition, guided by Differential Evolution algorithms and user feedback, in their case study resulted in six songs that were contracted during the experiment. Based on their research, they argued for human-driven evolutionary methods to maintain creativity and not be diminished solely by static datasets.

NLP-inspired architectures have also applied to AI's symbolic music generation. The text-based melody model MT-GPT-2, introduced by [16], was evaluated using the MEM method. Regarding the trade-off between reconstruction fidelity and attribute independence across genres, authors of [17] compared measureVAE and adversarialVAE. However, these studies point to the absence of standardized evaluation, in the same manner as [18], who noticed the lack of consolidation in model benchmarking and task taxonomy.

Other contributions include the exploration of the LSTM, Leak-GAN, and Music SketchNet models [16], which have achieved different representational designs and evaluations. Parallel concerns regarding dataset bias and minority genre representation [19] reflect the need for interpretable and inclusive AI models [17]. However, computational efficiency, real-time responsiveness, questions about output, and all forms of bias are among the challenges that practical deployment still grapples with [20].

2.1 Novelty of the Proposed Work

Although significant advances have been made in generating music with AI, most of the past literature has focused on individual tasks, such as melody-only generation [1], [2], lyrics-to-melody generation [4], or melody-to-lyrics generation [10], without presenting an integrated or interactive pipeline that enables real-time multi-modal control. Moreover, currently available methods based on GANs, Transformers, or rule-based systems are usually limited to interpretability, symbolic-to-audio correspondence, and semantic consistency between lyrics and melody [6], [7], [14].

To address these shortcomings, MelodyCraft proposed a single, modular, and interactive approach to synchronized lyrics-to-melody generation. This is novel not only in the sense that individual deep learning models are used, but also in the manner in which all these elements are bound together and trained to provide a bidirectional correspondence and real-time flexibility.

MelodyCraft has made many unique contributions, which can be summarized as follows:

- Two-Way Lyric-Melody Alignment: MelodyCraft is an integrated model of lyrical phrasing and melodic rhythm that ensures the presence of semantic, temporal, and emotional consistency between text and music, which is not often considered in parallel in other systems.
- Metadata-Conscious Conditioning Layer: The conditioning variables, genre, and emotion tags are integrated into the metadata of both the text and audio modules, which leads to form and sentiment-in-line with the compositions.
- Fast real-time control: Users can control parameters, including tempo, emotion, and duration, during inference by dynamically adjusting the parameters to produce controlled, interactive, and context-aware music.
- Logically Temporally Coherent Encoding Token Fusion: This system uses a common tokenization and decoding mechanism between textual embeddings and audio tokens to achieve structural and rhythmic alignment between lyrics and melody.

Thus, the innovativeness of MelodyCraft lies in its integrative design philosophy, which converts the traditionally discontinuous process of creating lyrics and melodies into a coherent, clear, and flexible, multimodal workflow. Therefore, this research-centered system focuses on the interpretability, controllability, and reproducibility of AI-assisted musical compositions.

3. METHODOLOGY

The MelodyCraft platform is a holistic and modular pipeline that is in harmony with datasets, deep learning, and an end-to-end interface to assist in the final music creation based on AI. This section summarizes the methodological basis and architecture of the unified process, which describes how the user inputs are converted into expressive and genre-relevant musical compositions.

3.1 Dataset Framework

MelodyCraft uses two major datasets to generate high-quality emotionally coherent songs because of the two-fold functional structure of the system: music composition and lyrics generation.

In audio synthesis, MelodyCraft combines MusicGen-small by Facebook AI Research (FAIR) [21], a transformer-based model that is pretrained on a large-scale dataset of more than 20,000 hours of professionally curated audio. The genres, emotional shades, instruments, rhythms, and tempos of this corpus are all diverse. The variety and richness of the dataset enabled the model to create high-fidelity, style-sensitive musical tracks for user-specified prompts. As a pre-trained model, MusicGen also removes the computational cost of training a model, which provides the system with fast deployment, responsiveness, and scalable integration into cloud-based services.

In addition, the lyrics generation component of MelodyCraft is based on a finely tuned variant of the Mixtral language model that is finely tuned to provide stylistically faithful and sentiment-aligned lyric generation. Domain-specific training was performed using the music_lyrics dataset.json, which is a corpus of lyrical data records in JSON format. The fields in the dataset are listed in Table 1.

This is systematized by genre and emotion conditioning, which allows the model to generate lyrics that are stylistically realistic, semantically coherent and resonant. Both expressive output and efficient inference are achieved by the sparse mixture-of-experts (MoE) architecture of Mixtral, which makes the module suitable for facilitating a real-time music generation workflow.

Table 1: *Structured fields of the music_lyrics.json dataset.*

Field	Example Value	Description
genres	[“canadian pop”, “pop”, “post-teen pop”]	List of genres associated with the song/track.
progress	“1.0”	Version or status indicator for the entry.
track_id	“0”	Internal identifier for the track.
start_lyric	“We were inseparable. Everything I had to do I did it next to you”	first line of the song’s lyrics used as a seed input.
processed _lyric	“We were inseparable. Then our love was interrupted by my schedule. You get used to being alone.”	Full lyric text used for training and model prediction.

3.2 Data Preprocessing and Prompt Engineering

Although MusicGen is already pretrained, the quality of the input has a significant impact on the output. MelodyCraft also implements a prompt engineering pipeline that enhances textual input provided by the user through lexical normalization, grammar correction, and semantic tagging. Attributes such as genre, mood, tempo, and instrument preferences can be specified by users, which are then run through the parsing and normalization process to align with the internal representations of the model.

The preprocessing of the lyric generation component is as follows:

- The data were cleaned to eliminate HTML tags, emoticons, special characters, and wasted white spaces.
- The lyrics were tokenized using byte pair encoding (BPE) to divide the text into subword units that are easy to manage.
- Labeling every record according to genre-emotion mappings to enable sampling conditional and stylistic anchoring.

In addition, the MelodyCraft interface allows the user to define the required audio length, 5-60 seconds, with respect to an interactive slider. This value is a smooth-moving parameter of the MusicGen generation pipeline, whereby the system can adjust the structural and temporal characteristics of the output using the same parameter. Fig. 1 illustrates the architectural flow of the Mixtral-based lyrics generation module, which consists of prompt conditioning, tokenization, and sparse expert routing.

3.3 Mixtral-based lyrics Generation Pipeline

A flowchart of the Mixtral-based lyrics generation module adopted by MelodyCraft is shown in Fig. 1. It starts with user-friendly input queries that comprise genre, mood, and context. The prompts were processed by cleaning, tokenizing and positional encoding. The tokens obtained were embedded with tokens and positional embedding layers to create a merged representation.

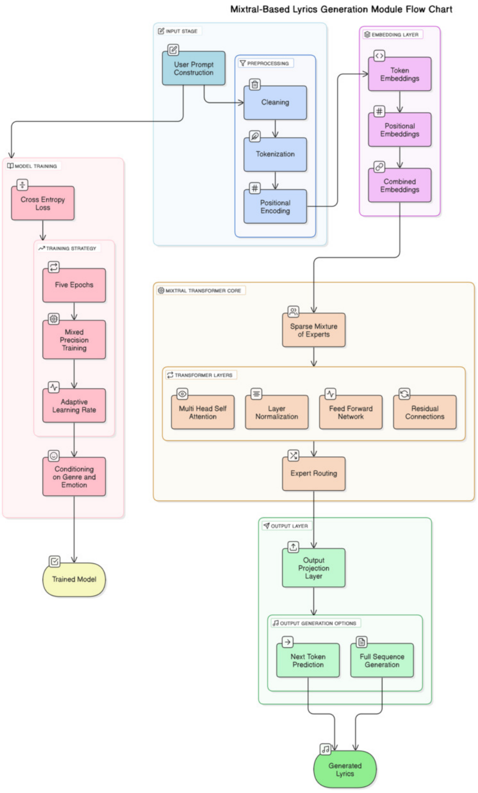


Fig.1: *Detailed architecture of the Mixtral-based lyrics generation pipeline, showing input tokenization, transformer blocks, and output projection.*

A sparse mixture-of-expert transformer architecture was then applied to this representation. Multi-head self-attention, layer normalization, feedforward networks, and residual connections are used within every transformer layer to learn deep contextual dependency. The expert routing process is a dynamically chosen mechanism that selects the appropriate subnetworks and allows for efficient and expressive lyric synthesis.

The training pipeline was highly effective in optimizing the model using cross-entropy loss, mixed-precision training, adaptive learning rates, and genre-emotion conditioning over five epochs. The transformer layer output is then reduced to an output layer projection and decoded by either next-token prediction or full-sequence generation, which eventually produces high-quality, contextually aware lyrics.

3.3.1 Fine-Tuning Configuration and Training Context

To maximize the lyric generation technology of the Mixtral module of MelodyCraft, a QLoRA-based 4-bit quantization approach was adopted to fine-tune the Mistral-7B-Instruct-v0.1 model. This method allowed for the efficient adaptation of the model with low memory and without sacrificing the model performance. Training was performed on an Nvidia A100 GPU (40GB VRAM) using the Hugging Face Transformers and Accelerate libraries.

Fine-tuning involved using a custom lyric generation corpus with approximately 28000 prompt-lyric pairs of various genres–pop, jazz, rock, lo-fi, hip-hop, and country–for diversity of style. The dataset was split into training, validation, and testing subsets in a ratio of 80:10:10 to balance the generalization and reliability of the evaluations.

Fine-tuning was performed with the AdamW optimizer (learning rate = 2*10-4) using a cosine decay scheduler with a 0.03 warm-up ratio, batch size = 16 (gradient accumulation = 4), weight decay = 0.01, and training in mixed-precision (FP16) for computational efficiency. The process was run for three epochs and took approximately eight hours on the A100 GPU. During adaptation, only the language modeling head and adapter layers were trainable, whereas the other core transformer layers of the base model were frozen to retain the pretrained linguistic knowledge. The random seed was set to 42 to ensure reproducibility of the results.

The fine-tuned model obtained a validation loss of approximately 1.85 and an estimated BLEU score of 0.65, confirming high lyrical coherence and stylistic fidelity to the genre after the training.

3.3.2 Model Architecture and Hyperparameter Specification

The lyric generation part is also based on the Mistral-7B-Instruct-v0.1 transformer architecture with approximately seven billion parameters, 32 transformer layers, 32 attention heads, and a hidden dimension of 4096. The base checkpoint (mistralai/Mistral-7B-Instruct-v0.1) was retrieved from the Hugging Face Hub.

Fine-tuning was performed using QLoRA adapters inserted into the attention and feed-forward blocks. Only the adapter parameters and the last language model head were updated, and the other backbone layers were frozen. The selective adaptation method reduced the computational cost and memory usage while preserving the rich, pretrained linguistic knowledge of the model. The configuration of the fine-tuning process, including the optimizer setup, learning schedule, batch size, and hardware specifications, is presented in Table 2.

Table 2: *Fine-tuning configuration, model architecture, and hyperparameter specifications for the Mixtral-based lyric generation module.*

Hyperparameter	Value / Setting
Optimizer	AdamW
Learning Rate	2×10^{-4} (cosine decay scheduler)
Batch Size	16 (gradient accumulation = 4)
Epochs	3
Precision	FP16 (mixed)
Weight Decay	0.01
Warm-up Ratio	0.03
Random Seed	42
Hardware	1 × NVIDIA A100 GPU (40 GB VRAM)
Total Compute Time	≈ 8 hours
Validation Loss	≈ 1.85
BLEU Score	≈ 0.65

3.4 MusicGen Model Architecture

The MusicGen-Small model is the foundation of the audio generation in MelodyCraft. A text encoder first a text encoder that maps textual prompts to high-dimensional embeddings. They are encoded into semantic representations using stacked layers of transformer decoders that encode the semantic representations into a series of discrete audio tokens. The token stream is fed into Encodec, an audio codec that restores the waveform using convolutional decoders.

The top-k sampling was set to 120, and the temperature was set to 0.8 to balance the diversity and coherence of audio. These parameters provide directed randomness because MelodyCraft can create diverse, musically meaningful audio songs of different genres and themes.

3.5 Integrated Backend and User Interface

It has a user experience enhanced by a responsive and intuitive front-end interface linked to a modular backend engine. The front-end interface permits users to:

Input: Song information: song title, genre, mood, tempo, style of lyrics, etc.

- The length of the song was changed using a real-time slider.
- Composition generation and playback
- Export audio files in MP3, WAV, or AIFF formats.
- Preview and editor Lyrics preview and edit the generated content using the lyrics preview and editor module.

The following modular components were used to create the backend system.

- NLP Processing Unit: This is an NLP library that uses NLP libraries (Hugging, spaCy, and NLTK) to interpret user intent and contextual parameters.

- Lyrics Module: Lyrics are created using a narrowly focused Mixtral transformer model, which substitutes the previous GPT-2 based prototype. This release was developed to improve contextual understandability, stylistic faithfulness, and emotion-conscious lyric synthesis.
- Melody Module: The semantic input is transformed into instrumental music with MusicGen.
- Voice Module: Imposes TTS and singing voice generation to convert vocal lines into lyrics.
- Audio Mixing Module: Aligns waves between vocals and instrumentals by processing time-aligned waveforms.
- Database Module: User sessions, historical prompt and generated assets.
- Export Module: Converts the finished audio into several distribution formats.

Fig. 2 shows the complete interaction between the front-end and back-end components at the system level.

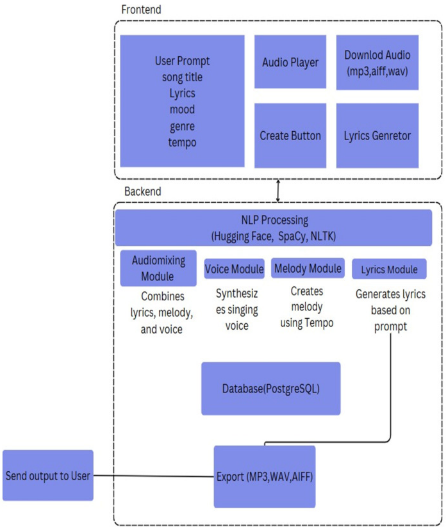


Fig.2: *High-level architecture of the MelodyCraft platform, showing the flow of user input through frontend components, backend processing units, and final audio output generation.*

3.5.1 Voice and Audio Synchronization Modules

The Voice Module in MelodyCraft is responsible for converting text-based lyrics into natural, expressive singing voices. This was performed within the framework of OpenVPI DiffSinger, a diffusion-probabilistic model tailored for singing voice synthesis. DiffSinger is a high-quality singing vocal synthesizer that conditions the denoising diffusion process for linguistic and musical features. In this process, the lyric text is first transformed into sequences of phonemes using a phonetic front-end. These phonemes are matched to the corresponding melody in the form of pitch and duration sequences (usually taken from a MIDI file or a symbolic score). These features were then encoded by the DiffSanger model, and the resulting output mel-spectrogram was refined by a diffusion-based decoder that progressively removed noise from the latent representation to provide smooth pitch transitions and expressive phrasing. Finally, the generated mel-spectrogram was transformed into an audio waveform using a HiFi-GAN vocoder, resulting in a lifelike singing voice with natural human intonation, timbre, and vibrato. The model allows explicit control of both pitch and duration, which can be used to accurately guide the coincidences between lyrical syllables and musical notes. This ensures that each phoneme is sung correctly in terms of timing and pitch, resulting in correct vocal and musical outputs.

Once the vocal track is synthesized, MelodyCraft uses a dedicated Audio Mixing Module to match the synthesized singing voice with the instrumental accompaniment. This process aligns two audio sources to ensure that they are rhythmically precise and harmonized between the two sources. The beat and onset detection functions applied to the instrumental track initiate the synchronization pipeline, where rhythmic anchor-point locations are identified and used to synchronize the instrumental tracks. Concurrent with the synthesis of the vocal waveform, syllabic onsets were extracted from the synthetic vocal waveforms using amplitude envelope analysis. A cross-correlation and dynamic time warping (DTW) algorithm was then used to compare the two onset patterns, such that the vocal phrases began and ended temporally in synchrony with the corresponding instrumental beats.

To maintain harmonic alignment, fundamental frequency (F0) analysis was performed on both vocal and instrumental tracks using short-time Fourier transform (STFT)-based spectral estimation. Local pitch correction is used whenever differences between the pitches of the vocal and accompaniment are detected using harmonic matching filters to maintain tonal coherence in the mix. For the final waveform-level integration, the system uses the PyDub audio processing library, which allows for accurate time-domain operations for waveform alignment, normalization, and mixing. The vocal and instrumental recordings were resampled to a universal sampling rate (44.1 kHz) and brought to an equal level of loudness and time spread according to the calculation of the onset offsets. The resulting mix was encoded into AudioSegment using PyDub, and gain adjustment, loudness normalization, and additive overlay blending were applied. Spectral balancing and dynamic range compression were then applied to the output to ensure consistent volume levels, tonal clarity, and a natural-sounding and professional mix between the synthesized vocals and instrumental accompaniment.

3.6 Token-Based Music Generation Workflow

Fig. 3 illustrates the entire process of composing music based on natural language input. It begins with the user input, which is then converted into dense embeddings by the text encoder. When the system was set to audio conditioning, the input waves were converted into discrete tokens using the Encodec WAV Encoder. Text-based embeddings combined with these tokens are then input into the MusicGen transformer decoder, which generates a new sequence of audio tokens.

The tokens were reverse-coded into audio using an Encodec Token Decoder, producing a high-fidelity waveform output. This procedure allows for the proper maintenance of musical structure, time, and tone.

MelodyCraft allows the creation of music in a scalable, customizable, and emotionally expressive manner through a balance between deep learning models and curated datasets, as well as modular system design. It is an integrated architecture that is not only computationally efficient but also has a high level of creative control by end users. Fig. 3 illustrates the token-based generation and decoding processes of MusicGen and Encodec.

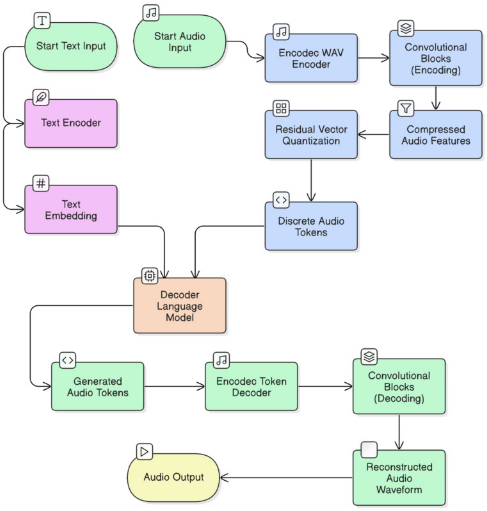


Fig.3: *Token-based music generation pipeline showing how text and audio inputs are encoded, decoded, and transformed into high-fidelity musical embedding.*

3.7 Comparative Spectrogram and Statistical Analysis of Human v/s AI-Generated Music

A comparative statistical study was performed on the spectral properties of MelodyCraft-generated compositions against human compositions to obtain a quantitative measure of the acoustic similarity between the compositions produced by the MelodyCraft algorithm and those composed by humans. This was performed using a benchmark dataset of 20 audio tracks, which included 10 non-professional human-composed audio tracks and 10 AI-generated tracks created using the MelodyCraft framework. To ensure generalizability, the dataset consisted of samples from three genres: pop, rock, and jazz, which also had a variety of rhythmic and harmonic characteristics.

Three important spectral descriptors were calculated for each track using the Librosa digital signal processing library in Python:

- Spectral Centroid: It center of mass of the sound spectrum, which is perceived in the brightness or timbre of the music.
- Spectral Bandwidth: This indicates the width of the spectral distribution around its center and is an indicator of the harmonic richness and complexity of the timbre.
- Spectral Roll-off: Refers to the frequency at which 85% of the spectral energy is contained and is usually used to differentiate between tonal and noisy signals.

The values of the features were computed for each file and grouped (human versus AI). Table 3 presents the values extracted from all 20 samples used in this analysis. The findings were input into the following statistical tests. Feature extraction was followed by a statistical comparison of the features using Welch's two-sample t-test, which considers unequal variances between groups. The level of significance was set at $\alpha = 0.05$. In addition to the numerical analysis, boxplots were generated as a visual presentation of the distribution and overlap of spectral features in both human- and AI-generated samples to provide an intuitive interpretation of the comparative behavior of the two sets of samples.

Table 3: *Examples of user-defined prompts and corresponding model responses.*

Music File Generated From	**centroid**	**bandwidth**	**rolloff**
human	1767.979	2254.272	3628.741
human	1949.457	2410.149	4318.295
human	2132.200	2450.933	4741.946
human	2749.177	2643.742	5692.661
human	2327.238	2591.402	5091.022
human	1304.501	1524.313	2563.775
human	2354.182	2660.297	5447.410
human	2537.689	2714.579	5811.522
human	1820.985	1863.960	3593.846
human	1719.121	2181.081	3704.187
AI	2556.651	2744.676	5352.418
AI	2630.189	2422.414	4854.467
AI	2374.747	2583.518	4951.458
AI	2564.541	2672.503	5721.097
AI	2704.800	2550.449	5283.759
AI	1683.201	2343.605	3495.363
AI	2122.717	2409.151	4290.930
AI	1850.354	1884.096	3340.804
AI	1741.389	2216.611	3573.804
AI	3078.229	2801.441	6385.772

3.8 Human Evaluation Study Design

To supplement the quantitative analysis of audio and fulfil the requirement of perceptual validation, a human test was conducted to determine the musical and emotional coherence of MelodyCraft-generated musical compositions in comparison with professionally composed musical pieces.

One hundred and sixty respondents (both musicians and non-musicians) were approached to appraise a collection of 18 anonymized audio recordings with an equal number of human-composed and AI-generated music pieces.

All participants wore headphones, listened to the clips in non-sequencing order, and rated them using a five-point Likert scale (1 = low, 5 = high).

The assessment was performed based on the following five major perceptual dimensions:

Melody Quality: The general good feeling and tonal attractiveness of the melody.

Lyrical Coherence with Melody: subjective congruence between words and music phrasing.

Emotional Generality to the Genre: The degree to which the composition conveyed the tone or feeling it was supposed to convey.

Rhythm and Flow: Temporal consistency and rhythmic naturalness.

Overall Creativity: A sense of originality and associative artistry of the work.

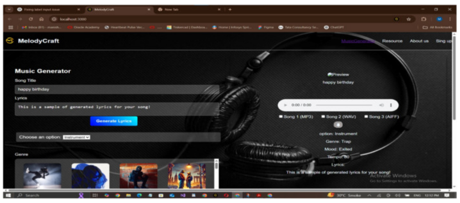
To guard against bias, the study was conducted anonymously, and all scores were normalized prior to aggregation. The mean and standard deviation were calculated for the human-composed and AI-generated tracks. All responses were normalized and aggregated to compute the Mean Opinion Scores (MOS) and standard deviations for each criterion, separately for human and AI-generated samples.

4. RESULT AND ANALYSIS

4.1 User Interface and System Interaction

The MelodyCraft platform has a user-friendly graphical user interface (GUI) that allows easy end-to-end music generation conveniently and easily. The interface enables users to input prompts in terms of song title, genre, mood, and type of lyrics, and a slider to regulate the required composition period (5–60 s). The interactive dashboard is presented in Fig. 4, and its main characteristics include timely input fields, audio preview in real time, and a download button to save the created composition for future use.

In terms of usability, the system was highly responsive and had low latency when synthesizing audio and generating lyrics. Other functions, such as token generation, prompt conditioning, and waveform rendering, are separated so that the user does not need to understand how they function. Therefore, they can operate seamlessly, even when the user is not technically sound. The usefulness of this interactive loop can be demonstrated by the fact that the system boasts real-time inference functionality, whereby the synthesis of lyrics and melody modules is integrated into a single output. The synthesis of NLP-prompt normalization and emotion tagging ensures that the interface is consistent with the corresponding logic of backend generation. Fig. 5 shows a sample of user input and system-generated responses, demonstrating the correspondence between prompts and outputs in the content and style of music.



Song Title: happy birthday

Lyrics: This is a sample of generated lyrics for your song!

Choose an option: Instrument

happy birthday

0:00 / 0:00

option: Instrument

Genre: Trap

Mood: Exited

Tempo: 80

Lyrics:

This is a sample of generated lyrics for your song!

Fig.4: *Graphical User Interface (GUI) of MelodyCraft, enabling user-friendly input of song attributes (genre, mood, lyrics type, duration) and providing real-time preview, editing, and export functionality.*

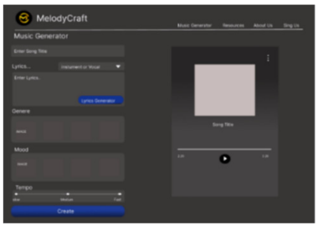


Fig.5: *Illustration of user-defined textual prompt and corresponding audio-lyric output generated by the MelodyCraft platform. This example demonstrates alignment between lyrical tone and musical structure.*

4.2 Audio Generation from Natural Language Prompts

The quality and relevance of the audio output of the MusicGen model were evaluated using various user-defined prompts that differed in terms of genre, mood, tempo, and form. Table 4 shows the sample prompts and settings for their generation, such as acoustic ballads, energetic pop songs, and melancholic jazz instruments.

MelodyCraft translates semantic prompts into high-fidelity audio files with the required emotional tone and music genre. An example is the prompt of a sad acoustic ballad about lost love, which produces a slow, melancholic melody with minimal instrumentation. An energetic pop song is a good summer vibe prompt that produces an upbeat and rhythmic melody that is typical of modern pop frameworks.

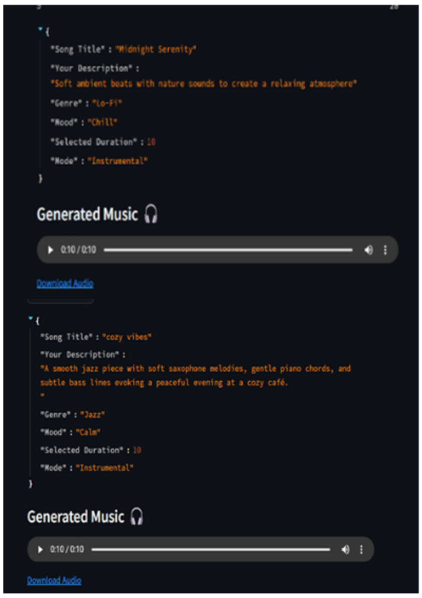
MelodyCraft translates semantic prompts into high-fidelity audio files with the required emotional tone and music genre. An example is the prompt of a sad acoustic ballad about lost love, which produces a

slow, melancholic melody with minimal instrumentation. An energetic pop song is a good summer vibe prompt that produces an upbeat and rhythmic melody that is typical of modern pop frameworks.

Table 4: *Examples of user-defined prompts and corresponding model responses.*

Prompt Example	Genre	Mood	Tempo	Output Type
"A sad acoustic ballad about lost love"	Acoustic	Sad	Slow	Full Song
"Energetic pop song for summer vibes"	Pop	Happy	Fast	30-second Clip
"Melancholic jazz tune with saxophone solo"	Jazz	Melancholic	Medium	Instrumental Only

Controlled variability was added using top-k sampling (top-k was set to 120) and a temperature value (0.8), which enabled the production of unique but coherent compositions. This was especially noticeable in prompts that required mood-specific instrumentation (e.g., jazz solos, or ambient synths). The findings confirm that the MusicGen model, in combination with token-based conditioning and high-resolution audio reconstruction of Encodec, generates stylistically diverse and sonically similar outputs to those of human-generated music. The output-generated music responses for the prompt-specific stimuli are illustrated in Fig. 6.



```
{
  "Song Title" : "Midnight Serenity"
  "Your Description" :
  "Soft ambient beats with nature sounds to create a relaxing atmosphere"
  "Genre" : "Lo-Fi"
  "Mood" : "Chill"
  "Selected Duration" : 10
  "Mode" : "Instrumental"
}
```

Generated Music 🎧

Download Audio

```
{
  "Song Title" : "cozy vibes"
  "Your Description" :
  "A smooth jazz piece with soft saxophone melodies, gentle piano chords, and
  subtle bass lines evoking a peaceful evening at a cozy café.
  "
  "Genre" : "Jazz"
  "Mood" : "Calm"
  "Selected Duration" : 10
  "Mode" : "Instrumental"
}
```

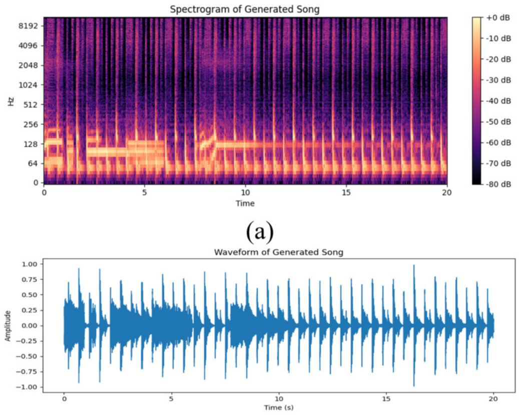
Generated Music 🎧

Download Audio

Fig. 6: *Illustration of user-defined textual prompt and corresponding audio-lyric output generated by the MelodyCraft platform. This example demonstrates alignment between lyrical tone and musical structure.*

4.2.1 Visual Analysis of Generated Audio

Waveform and mel-spectrogram visualizations of a series of output tracks were performed to measure the quality, structure, and genre consistency of the generated music (Figs. 7a-9b). Fig. 7(a) depicts a pop-style track with regular rhythmic peaks and amplitude modulations, indicating an estimated tempo of approximately 120 beats per minute (BPM). The spectrogram in Fig. 7(b) indicates a significant energy distribution in both the low (64–512 Hz) and high-frequency bands (2–8 kHz), suggesting a layered composition with bass, drums, and synthetic harmonics, which is typical of modern pop music.



(a)

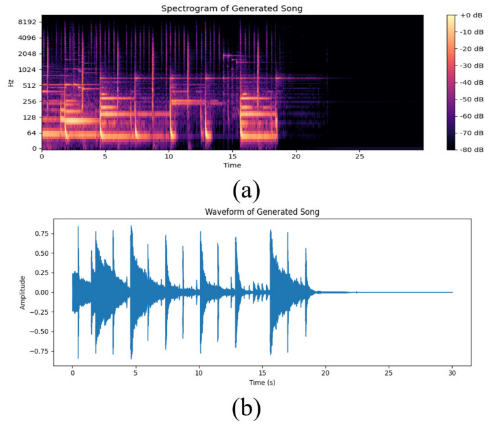
(b)

Fig. 7: *(a) Waveform of the generated pop track showing regular amplitude peaks and rhythmic consistency, suggesting a tempo of approximately 120 BPM. (b) Mel-spectrogram of the same pop track displaying strong energy in both the low (64–512 Hz) and high (2–8 kHz) frequency ranges, indicating layered instrumentation.*

In comparison, Fig. 8(a) and 8(b) show a low-fidelity or cinematic output. Fig. 8a shows the waveform-decaying amplitude envelopes and the sparsity of rhythmic pulses, whereas Fig. 8(b) indicates low-frequency energy with little high-end activity. This emphasizes the flexibility of the model in terms of time semantics through the generation of ambient and relaxed textures with sparse percussions.

In Figs. 9(a) and 9(b), the output of a rhythm-oriented track can be observed, and it probably fits best in genres such as EDM and hip-hop. Fig. 9a reveals that the regularity of the amplitude peaks in the waveform is extremely high over the 30-second duration, which is evidence of high beat alignment. The spectrogram (Fig. 9b) is characterized by a vertically constant frequency activation over time and spectrum (64–2048 Hz) and confirms a beat-structured structure with a harmonic overlay.

Combined, these visualizations indicate that MelodyCraft can create stylistically correct audio that reflects the rhythmic structure and spectral dis-

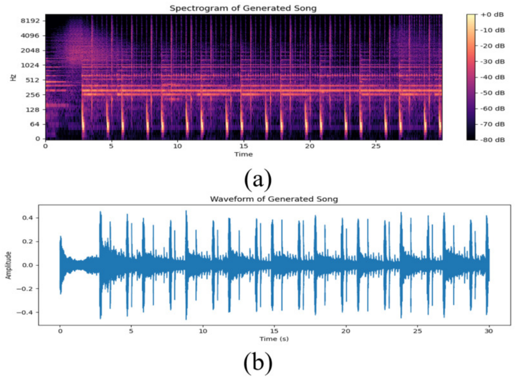


(a)

(b)

Fig.8: *(a) Waveform of a generated lo-fi or cinematic track, characterized by a gradually decaying amplitude and low rhythmic density. (b) Mel-spectrogram of the same track showing dominant low-frequency bands and minimal upper-harmonic content, which is typical of ambient music.*

tribution, which are held together in a coherent and genre-aware manner.



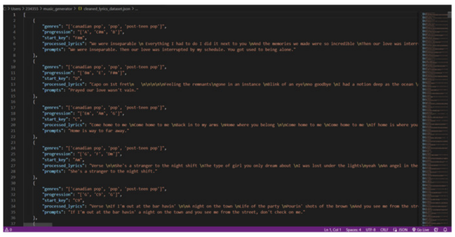
(a)

(b)

Fig.9: *(a) Waveform of rhythm-focused track (hip-hop), marked by evenly spaced amplitude spikes reflecting strong beat alignment. (b) Mel spectrogram of the same track showing vertically structured frequency activation across low-to-mid bands, consistent with percussive and harmonic layering.*

4.3 Lyric Generation Results

The performance of the lyric generation was analyzed using a set of curated prompts that were marked with genre and emotion tags. A snapshot of the dataset is presented in Fig. 10, which shows the contribution of genre-emotion conditioning to the generation of stylistically appropriate lyrics. The input to the Mixtral-based lyrics generation module shown in Fig. 11 was as follows: "A heartfelt rock ballad about overcoming failure. The ensuing lyrics embodied the storytelling aspects of struggle, self-redemption, and hope, which aligned with the thematic tendencies of the rock genre. The model also exploited a bare-bone mixture-of-experts structure used to dynamically route the inputs to a set of specialized transformer subnetworks, which also led to expressiveness and contextual coherence.

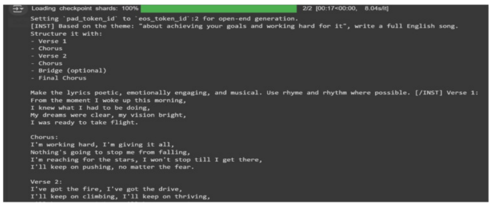

C: > Users > 234355 > music_generator > {} cleaned_lyrics_dataset.json > ...
Ln 1, Col 1   Spaces: 4   UTF-8   CRLF   {} JSON   Go Live

```
[
    {
        "genres": "['canadian pop', 'pop', 'post-teen pop']",
        "progression": "['A', 'C#m', 'B']",
        "start_key": "F#m",
        "processed_lyrics": "We were inseparable \n Everything I had to do I did it next to you \nAnd the memories we made were so incredible \nThen our love was interr
        "prompts": "We were inseparable. Then our love was interrupted by my schedule. You got used to being alone."
    },
    {
        "genres": "['canadian pop', 'pop', 'post-teen pop']",
        "progression": "['Bm', 'E', 'F#m']",
        "start_key": "D",
        "processed_lyrics": "Capo on 1st fret\n   \n\n\n\n\nFeeling the remnants\ngone in an instance \nBlink of an eye\nno goodbye \nI had a notion deep as the ocean \
        "prompts": "Prayed our love wasn't vain."
    },
    {
        "genres": "['canadian pop', 'pop', 'post-teen pop']",
        "progression": "['Em', 'Am', 'G']",
        "start_key": "C",
        "processed_lyrics": "Come home to me \nCome home to me \nBack in to my arms \nHome where you belong \n\nCome home to me \nCome home to me \nIf home is where you
        "prompts": "Home is way to far away."
    },
    {
        "genres": "['canadian pop', 'pop', 'post-teen pop']",
        "progression": "['G', 'F', 'Dm']",
        "start_key": "Am",
        "processed_lyrics": "Verse \n\nShe's a stranger to the night shift \nThe type of girl you only dream about \nI was lost under the lights\nyeah \nAn angel in the
        "prompts": "She's a stranger to the night shift."
    },
    {
        "genres": "['canadian pop', 'pop', 'post-teen pop']",
        "progression": "['G', 'C9', 'G']",
        "start_key": "C9",
        "processed_lyrics": "Verse \nIf I'm out at the bar havin' \n\nA night on the town \nLife of the party \nPourin' shots of the brown \nAnd you see me from the str
        "prompts": "If I'm out at the bar havin' a night on the town and you see me from the street, don't check on me."
    },
    {
```

Fig.10: *Snapshot of the structured music_lyrics.json dataset used to train the Mixtral model. Each entry includes genre, emotion tags, start lyrics, processed lyrics, and metadata for conditioning and model tuning.*

The evaluation metrics proved that the model had good performance: its cross-entropy loss stopped decreasing after five epochs, and the BLEU scores ranged between 0.5 and 0.7 for different genres. This means that there is a high level of textual fluency and semantic relevance compared to the training corpus. Fig. 12 depicts the structural alignment, as there was a correct progression of the verses and emotional flow in the lyrics created. These findings justify the efficacy of genre-emotion-conditioned prompts and sparse expert routing in generating high-quality lyrical content.

Prompt: "A heartfelt rock ballad about overcoming failure' Generated Lyrics (excerpt)":



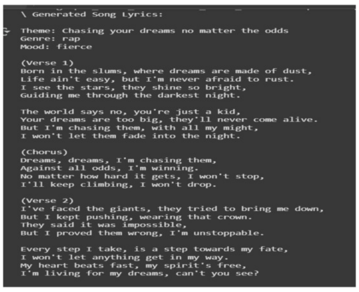
```
Loading checkpoint shards: 100%                                  2/2 [00:17<00:00,  8.04s/it]
Setting `pad_token_id` to `eos_token_id`:2 for open-end generation.
[INST] Based on the theme: "about achieving your goals and working hard for it", write a full English song.
Structure it with:
- Verse 1
- Chorus
- Verse 2
- Chorus
- Bridge (optional)
- Final Chorus

Make the lyrics poetic, emotionally engaging, and musical. Use rhyme and rhythm where possible. [/INST] Verse 1:
From the moment I woke up this morning,
I knew what I had to be doing,
My dreams were clear, my vision bright,
I was ready to take flight.

Chorus:
I'm working hard, I'm giving it all,
Nothing's going to stop me from falling,
I'm reaching for the stars, I won't stop till I get there,
I'll keep on pushing, no matter the fear.

Verse 2:
I've got the fire, I've got the drive,
I'll keep on climbing, I'll keep on thriving,
```

Fig.11: *Generated lyrics excerpt from the Mixtral-based module in response to the prompt "A heartfelt rock ballad about overcoming failure." The lyrics demonstrate genre-specific narrative development and emotional arcs.*

4.4 Statistical Comparison of Spectral Characteristics

To further investigate the spectral fidelity of MelodyCraft-generated music, three quantitative characteristics, namely, spectral centroid, bandwidth, and roll-off, were contrasted with the characteristics of human-composed music using Welch's t-test. The resulting p-values (0.205, 0.386, and 0.579, respectively) showed that there were statistically significant differences in the mean spectra of the two groups at the $\alpha = 0.05$ level.

Although the AI-generated tracks were slightly higher in terms of all three features, the differences were insignificant, and the statistical significance was low. These findings imply that the spectral characteristics of MelodyCraft-generated music are similar to those of human music, particularly the energy of the frequency and timbral balance distribution.



```
\ Generated Song Lyrics:

Theme: Chasing your dreams no matter the odds
Genre: rap
Mood: fierce

(Verse 1)
Born in the slums, where dreams are made of dust,
Life ain't easy, but I'm never afraid to rust.
I see the stars, they shine so bright,
Guiding me through the darkest night.

The world says no, you're just a kid,
Your dreams are too big, they'll never come alive.
But I'm chasing them, with all my might,
I won't let them fade into the night.

(Chorus)
Dreams, dreams, I'm chasing them,
Against all odds, I'm winning.
No matter how hard it gets, I won't stop,
I'll keep climbing, I won't drop.

(Verse 2)
I've faced the giants, they tried to bring me down,
But I kept pushing, wearing that crown.
They said it was impossible,
But I proved them wrong, I'm unstoppable.

Every step I take, is a step towards my fate,
I won't let anything get in my way.
My heart beats fast, my spirit's free,
I'm living for my dreams, can't you see?
```

Fig.12: *Structural layout of the generated lyrics showing coherent verse progression, line-to-line consistency, and sentiment alignment. The output reflects the effectiveness of sparse expert routing and prompt conditioning.*

These findings are visually supported by the comparative boxplots in Fig. 13, which show a significant overlap between the two groups in all spectral dimensions. The statistical findings were supported by the interquartile ranges (IQRs), medians, and general distribution spreads, which demonstrated a significant spectral similarity between human and artificial intelligence-generated pieces of music. The computed statistics are presented in Table 5.

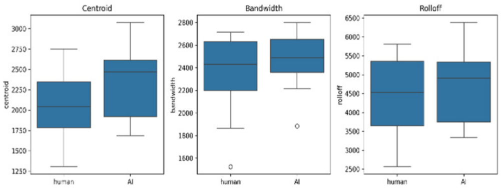


Fig.13: *Comparative boxplots illustrating the distribution of the spectral centroid, bandwidth, and roll-off for human-composed and MelodyCraft-generated music samples.*

The absence of statistically significant differences indicates that MelodyCraft produces audio with a spectral energy distribution comparable to that of human-created music. This implies that the system can reproduce basic acoustic attributes, such as brightness, harmonic unleash, and high-frequency roll-off, with a high level of realism.

The results of this comparative spectral analysis provide quantitative evidence that MelodyCraft-generated compositions exhibit acoustic properties consistent with those of professional human music. The following section presents a broader framework-level comparison of MelodyCraft with previous AI-based music-generation methodologies.

Table 5: *Welch's t-Test Results for Human vs. MelodyCraft-Generated Music.*

Feature	t-Statistic	p-Value	Human Mean	AI Mean
Spectral Centroid	–1.314	0.205	2066.25	2330.68
Spectral Bandwidth	–0.892	0.386	2329.47	2462.85
Spectral Roll-off	–0.565	0.579	4459.34	4724.99

4.5 Human Evaluation of Generated Music

A human-based listening study was conducted to compare human-created and MelodyCraft-created music across five perceptual dimensions. The ratings were averaged, and the mean results of all 16 participants are presented in Table 6. The aggregate statistics are shown in Table 7.

According to the results in Tables 6 and 7, the respondents rated human-composed and AI-generated music similarly across all perceptual dimensions. The AI-generated songs received slightly higher mean ratings for lyrical coherence, emotional alignment, and overall creativity, although the ratings for melody quality and rhythm were similar to those of human-generated songs. The small standard deviations (approximately 1.112) represent the similarity in the raters' opinions, which proves the accuracy of the assessment.

These findings indicate that MelodyCraft can generate music that audiences perceive as musically coherent, emotionally expressive, and artistically attractive. The similar MOS values in the two groups revealed that the system attained a perceptual quality similar to that of human composition.

The results confirmed that the multimodal architecture of MelodyCraft, which combines the processes of lyric writing, melody composition, and rhythmic composition, is effective in the quantitative and qualitative structures of human music composition. Such perceptual confirmation validates the synthesis capabilities of the framework in creating not only statistically accurate outputs but also aesthetically agreeable and creatively authentic in the eyes of human listeners.

Human evaluation confirmed that MelodyCraft achieves parity with professional compositions in both technical and perceptual domains. The following section provides a framework-level comparison of MelodyCraft with other state-of-the-art AI-based music generation approaches.

Table 6: *Mean Opinion Scores (MOS) of Human Evaluators for Individual Human- and AI-Generated Music Samples Across Five Perceptual Criteria.*

Music File Generated From / Criteria	**Melody Quality**	**Lyrical Coherence with Melody**	**Emotional Alignment with the Prompt-Genre**	**Rhythm and Flow**	**Overall Creativity**
Human	4.19	4.19	4.19	4.31	4.25
Human	3.88	3.81	3.69	4.06	3.94
Human	3.88	3.88	3.81	3.88	3.94
Human	3.88	3.25	3.75	3.81	3.88
Human	3.88	3.31	3.81	4.00	3.88
Human	4.13	4.00	3.50	3.69	3.94
Human	3.81	3.81	3.63	3.94	3.81
Human	3.88	3.63	3.56	4.00	3.94
Human	3.75	3.56	3.56	3.81	3.75
AI	3.94	3.88	3.94	3.94	4.19
AI	3.63	3.63	3.81	3.81	3.56
AI	4.13	4.13	4.19	4.06	4.25
AI	4.13	4.06	4.06	4.00	4.19
AI	4.06	4.06	4.00	3.88	3.94
AI	3.94	4.13	3.88	3.81	4.06
AI	3.88	4.13	4.00	4.00	3.94
AI	3.94	3.88	3.81	3.94	3.94
AI	3.88	3.81	3.75	3.88	3.88

Table 7: *Overall Mean Opinion Scores (MOS) for Human-Composed and AI-Generated Music.*

Evaluation Criterion	Human-Composed Music (Mean ± SD)	AI-Generated Music (Mean ± SD)
Melody Quality	3.91 ± 1.12	3.94 ± 1.17
Lyrical Coherence with Melody	3.71 ± 1.27	3.96 ± 1.19
Emotional Alignment with Prompt/Genre	3.72 ± 1.19	3.93 ± 1.18
Rhythm and Flow	3.94 ± 1.18	3.92 ± 1.18
Overall Creativity	3.92 ± 1.14	3.99 ± 1.17

4.6 Framework-Based Comparison of Music Generation Approaches

Because no standardized benchmarks have been established and no publicly accessible full-stack music generation systems are available, it is not possible to directly and quantitatively compare these models with previous models. Table 8 provides a comparative summary of the methodologies and reported results of the selected studies on AI-based music generation, including the proposed system. It is an organized source of methodological research without uniform assessment procedures.

Table 8: *Comparative Analysis of Methodologies and Outcomes in AI-based Music Generation Studies.*

Study	**Methodology Used**	**Key Outcome**
[1]	RNN, LSTM, Attention-based DL	Melodic compositions mimicking human style
[2]	RNN-GAN Hybrid	Expressive music using hybrid architecture
[3]	Transformer, EXPLING, Prior Attention	Controllable syllable-level lyric generation
[4]	LSTM, Memofu, RSE, SeqLoss	Style-aware melody generation
[5]	Transformer, Mutual Information	Improved interpretability via semantic alignment
[6]	Conditional GAN with Gumbel-Softmax	Interactive and consistent melody generation
[7]	TBC LSTM-GAN	High-quality attribute-specific melody generation
[8]	LSTM-GAN with RL	Coherent music via RL-integrated GANs
[10]	Rule-based AI Algorithm	Fully rule-based singable children's music
[11]	Hierarchical Gen. with Pretrained LM	Topic-aligned lyrics without aligned training data
[12]	Review of RNN, LSTM, MusicVAE	Broad applicability in education and therapy
[13]	Survey on DL-based Music Models	Comprehensive survey on DL techniques
[14]	Melody-guided Diffusion Model	Melody-conditioned diffusion-based music generation
[15]	DIWFA-GAN with GTO & Swish Activation	Emotionally expressive and coherent melodies
[16]	Differentia Evolution with Feedback	Commercially viable user-guided songs
[17]	MT-GPT2, MEM	Transfer learning for symbolic melody
[18]	MeasureVAE, AdversarialVAE	Improved structure and independence in music
[19]	Taxonomy Analysis	Framework classification and evaluation gaps
[20]	Dataset Bias Highlight	Bias and inclusivity challenges in datasets
Proposed System: Melody Craft	MusicGen + Mixtral (Transformer-based multimodal fusion)	Melody-lyrics alignment with adaptive fusion

5. CONCLUSION

This paper presents MelodyCraft, a unified system in which both lyrics generation and audio synthesis of instruments are incorporated into a single prompt-based music generation system. In contrast to more fragmented models that process individual music components, MelodyCraft converts textual and auditory modalities to generate complete musical compositions of the designated genre. The coupling of MusicGen-Small and a lyric generator built with Mixtral makes it possible to generate conditional music based on mood, genre, tempo, and thematic specifications (user inputs). It uses transformer-based systems, token-level decoding, and emotion-genre correspondence to produce consistent lyrical and musical texts. The semantic stability and quality of the generated lyrics were checked with the help of the BLEU scores, whereas the clarity and fidelity of the generated audio were checked with the help of spectrogram and waveform analysis. In the absence of standardized datasets and complete-stack assessment procedures, this study provides an overview of the comparative methodology to put the architecture of MelodyCraft into perspective with existing AI music-generation studies. In the future, MelodyCraft offers a basis on which future improvements can be made, including multilingual, adaptive sampling, and human-in-the-loop (HITL) editing. The interpretable and modular design of the system allows for scaling of research and creative exploration. Finally, MelodyCraft builds on the vision of AI-assisted music creation by providing a platform that is accessible, real-time, and controllable to enable creators in both the artistic and technical spheres.

Future work will address the development of a more comprehensive evaluation framework that includes objective perceptual metrics (such as the well-known Frrechet Audio Distance (FAD)) to establish a reference-free measure of audio quality and realism. For the lyrical part, semantic similarity metrics (e.g., BERTScore) allow one to better assess the textual relevance and emotional accuracy of the lyrics between the prompts and generated lyrics. In addition, a large-scale human listening study with a larger pool of listeners (≥ 30) will be performed to increase the statistical validity and inter-rater reliability of the perceptual evaluations. Ablative studies will also be conducted to understand the contribution of different critical building blocks, such as the metadata-aware conditioning layer, alignment mechanism, and encoding-based fusion to the overall system performance.

AUTHOR CONTRIBUTIONS

Conceptualization, A.C., R.J., H.V., K.K., and M.M.; Formal analysis, A.C., R.J., H.V., K.K., and M.M.; Methodology, A.C., R.J., H.V., K.K., and M.M.; Validation, A.C., R.J., H.V., K.K., and M.M.; Visualization, A.C., R.J., H.V., K.K., and M.M.; Writing – original draft, A.C., R.J., and H.V.; Writing – review & editing, R.J. All authors have read and agreed to the published version of the manuscript.

ETHICAL AND LICENSING COMPLIANCE

The Facebook/MusicGen-small model used in this study is licensed under the Creative Commons Attribution-Non-commercial 4.0 International (CC BY-NC 4.0) license. Accordingly, all uses of the model, including fine-tuning, inference, and dissemination of generated outputs, were strictly for academic and non-commercial research purposes. All research experiments and generated outputs were performed only for scholarly analysis and demonstration purposes, complying with the principles of fair academic usage and responsible AI research practice. No part of the audio or lyrics generated was used for redistribution, monetization, or any derivative commercial activity. Any potential commercial use or redistribution of MusicGen or any derivative thereof requires explicit licensing and authorization from copyright holders.

References

[1] M. Mane, "MelodyAI: AI-Powered Melodies Generation," *Interantional Journal Of Scientific Research In Engineering And Management*, vol. 7, no. 11, pp. 1–11, Nov. 2023.

[2] V. R. Bhaddurgatte and S. S, "Generating Music Using Machine Learning," *Interantional Journal of Scientific Research in Engineering and Management*, vol. 9, no. 1, pp. 1–9, Jan. 2025.

[3] Z. Zhang, Y. Yu and A. Takasu, "Controllable Syllable-Level Lyrics Generation From Melody With Prior Attention," in *IEEE Transactions on Multimedia*, vol. 26, pp. 11083-11094, 2024.

[4] Z. Zhang, Y. Yu and A. Takasu, "Controllable lyrics-to-melody generation," *Neural Computing and Applications*, vol. 35, no. 27, pp. 19805–19819, Sep. 2023.

[5] W. Duan, Y. Yu, X. Zhang, S. Tang, W. Li and K. Oyama, "Melody Generation from Lyrics with Local Interpretability," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 19, no. 3, pp. 1–21, May 2023.

[6] W. Duan, Z. Zhang, Y. Yu and K. Oyama, "Interpretable Melody Generation from Lyrics with Discrete-Valued Adversarial Training," in *Proceedings of the 30th ACM International Conference on Multimedia*, New York, NY, USA: ACM, pp. 6973–6975, Oct. 2022.

[7] A. Srivastava *et al.*, "Melody Generation from Lyrics Using Three Branch Conditional LSTM-GAN," in *MultiMedia Modeling.* MMM 2022.

Lecture Notes in Computer Science, pp. 569–581, 2022.

[8] M. Singhal, B. Saxena, A. P. Singh and A. Baranwal, "Study of the effectiveness of Generative Adversarial Networks towards Music Generation," in *2023 Second International Conference on Informatics (ICI)*, pp. 1–5, Nov. 2023.

[9] C. Liao, "AI-Algorithmically-Generated Song with Lyrics," in *2023 IEEE International Conference on Big Data (BigData)*, pp. 4495–4496, Dec. 2023.

[10] Y. Tian *et al.*, "Unsupervised Melody-to-Lyrics Generation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 9235–9254, 2023.

[11] Z. Chen, "Composing music under certain conditions based on neural network," *Applied and Computational Engineering*, vol. 64, no. 1, pp. 186–192, Jun. 2024.

[12] M. J. Pathariya, P. Basavraj Jalkote, A. M. Patil, A. Ashok Sutar and R. L. Ghule, "Tunes by Technology: A Comprehensive Survey of Music Generation Models," *2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS)*, Coimbatore, India, pp. 506-512, 2024.

[13] S. Wei, M. Wei, H. Wang, Y. Zhao, and G. Kou, "Melody-Guided Music Generation," *arXiv preprint* arXiv:2409.20196, 2024.

[14] T. Shaikh and A. Jadhav, "Music Generation Using Dual Interactive Wasserstein Fourier Acquisitive Generative Adversarial Network," *International Journal of Computational Intelligence and Applications*, vol. 24, no. 01, Mar. 2025.

[15] J. Kilb and C. Ellis, "Conserving Human Creativity with Evolutionary Generative Algorithms: A Case Study in Music Generation," *arXiv preprint* arXiv:2406.05873, 2024.

[16] Y. Guo, Y. Liu, T. Zhou, L. Xu and Q. Zhang, "An automatic music generation and evaluation method based on transfer learning," *PLoS One*, vol. 18, no. 5, p. e0283103, May 2023.

[17] N. Bryan-Kinns, B. Zhang, S. Zhao and B. Banar, "Exploring Variational Auto-encoder Architectures, Configurations, and Datasets for Generative Music Explainable AI," *Machine Intelligence Research*, vol. 21, no. 1, pp. 29–45, Feb. 2024.

[18] S. Ji, X. Yang and J. Luo, "A Survey on Deep Learning for Symbolic Music Generation: Representations, Algorithms, Evaluations, and Challenges," *ACM Comput Surveys*, vol. 56, no. 1, pp. 1–39, Jan. 2024.

[19] D. Wen, A. Soltan, E. Trucco and R. N. Matin, "From data to diagnosis: skin cancer image datasets for artificial intelligence," *Clin Exp Dermatol*, vol. 49, no. 7, pp. 675–685, Jun. 2024.

[20] D. Patil, N. L. Rane, P. Desai and J. Rane, "Machine learning and deep learning: Methods, techniques, applications, challenges, and future research opportunities," in *Trustworthy Artificial Intelligence in Industry and Society*, Deep Science Publishing, pp. 28-81, 2024.

[21] facebook/musicgen-small, "Hugging Face," Hugging Face. [Online]. Available: `https://huggingface.co/facebook/musicgen-small`

Ayushi Chauhan is an Information Technology graduate from Marwadi University, Rajkot. She possesses strong skills in Python, JavaScript, ReactJS, and modern web technologies. Her research interests include artificial intelligence, machine learning, natural language processing, and data-driven intelligent systems, with a keen focus on scalable and user-centric software solutions.

Rituraj Jain is an accomplished academician and researcher in Computer Science and Engineering with over 22 years of teaching and research experience. He is an Assistant Professor in the Department of Information Technology at Marwadi University, India. His research interests include AI, machine learning, deep learning, big data analytics, cyber-physical systems, and cloud computing. He has authored over 100 publications, contributed to IGI Global books, and holds multiple international patents in AI-driven healthcare and intelligent systems.

Harshalkumar Vanpariya is an Information Technology graduate from Marwadi University, Rajkot. He has strong skills in full-stack web development, including HTML, CSS, JavaScript, PHP, Python, and database management. His research interests include artificial intelligence, machine learning, computer vision, and intelligent human–computer interaction systems. He has published an IEEE paper on AI-driven mood and pose detection systems.

Keyur Kacha is an Information Technology graduate from Marwadi University, Rajkot. He has strong skills in full-stack web development, including HTML, CSS, JavaScript, PHP, Python, and database management. His research interests include artificial intelligence, machine learning, computer vision.

Manisha Makawana is an Information Technology graduate from Marwadi University. She possesses strong skills in data analytics and business intelligence, with expertise in Power BI, SQL, Excel, and data visualization. Her research interests include data-driven decision-making, business analytics, and intelligent reporting systems.