International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 36–48, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101005

# Intelligent Polyphonic Music Generation Using Large Language Models and Transformer-Based Sequence Learning

**[1]Puteri Amirah Ismail, [2]Muhammad Arif**

[1,2]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Malaysia

***Abstract*** **- Artificial Intelligence (AI) has significantly transformed the field of music generation by enabling machines to compose melodies, harmonies, and complete musical arrangements with minimal human intervention. Among various music generation tasks, polyphonic music generation remains one of the most challenging due to the simultaneous interaction of multiple musical voices, harmonic structures, rhythmic dependencies, and long-term compositional coherence. Traditional rule-based and statistical methods often struggle to capture the complexity of polyphonic compositions. Recent advancements in Large Language Models (LLMs), particularly Generative Pre-trained Transformer (GPT) architectures, have demonstrated exceptional capabilities in sequence modeling and creative content generation. This research investigates a GPT-based polyphonic music generation algorithm that leverages transformer architectures to learn musical representations from large-scale symbolic music datasets. The proposed framework utilizes MIDI-based music encoding, tokenization strategies, transformer-based sequence modeling, and autoregressive generation techniques to create coherent polyphonic musical compositions. Experimental analysis demonstrates that the GPT-based model effectively captures harmonic progression, rhythmic consistency, and melodic structure while generating musically plausible outputs. The study highlights the potential of large language models in computational creativity and intelligent music composition systems.**

***Keywords:*** Polyphonic Music Generation, GPT, Large Language Models, Transformer Architecture, Artificial Intelligence, Computational Creativity, MIDI, Deep Learning.

## 1. Introduction

Music is a universal form of artistic expression that combines melody, harmony, rhythm, and timbre to convey emotions and ideas. The process of music composition traditionally requires extensive musical knowledge, creativity, and experience. With the emergence of artificial intelligence, researchers have explored methods for automating music composition and assisting musicians in creative tasks. AI-based music generation systems have evolved from rule-based approaches to sophisticated deep learning architectures capable of generating highly realistic musical sequences [1].

Polyphonic music generation refers to the creation of musical compositions containing multiple simultaneous notes or voices. Unlike monophonic music, where only one melody line is present, polyphonic compositions involve complex interactions among melodies, harmonies, and rhythmic structures. Modeling these relationships requires understanding both local note dependencies and long-range musical patterns [2].

The development of Transformer-based Large Language Models such as GPT has revolutionized sequence generation tasks in natural language processing. These models employ self-attention mechanisms that effectively capture long-term dependencies within sequential data. Since symbolic music can be represented as sequences of musical events, GPT architectures offer promising opportunities for generating high-quality polyphonic music. This research investigates the application of GPT large models to polyphonic music generation and evaluates their effectiveness in producing coherent and aesthetically pleasing musical compositions [3].

Music has most expressive forms of human creativity, serving as a medium for communication, cultural preservation, and emotional expression. With the rapid automatic music generation has attracted significant attention from researchers, musicians, and the entertainment industry. Early computational music generation systems relied on predefined musical rules, probabilistic models, and handcrafted composition techniques, which were capable of generating only simple melodies with limited creativity and musical diversity. As music composition involves complex relationships among melody, harmony, rhythm, tempo, dynamics, and musical structure, conventional algorithms often struggle to capture these intricate

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 36–48, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101005

dependencies, particularly in where multiple notes and instruments. Consequently, there is an increasing demand for intelligent algorithms musical patterns directly from large-scale music datasets [4].

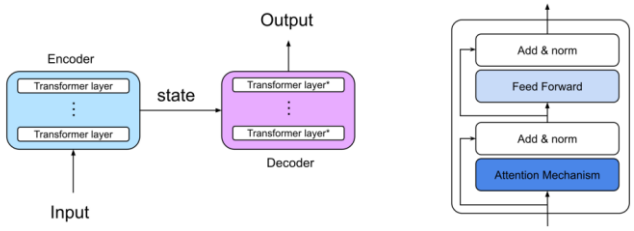



**Figure 1: Large Language Models (LLMs)**

This autoregressive learning framework enables GPT models to predict subsequent musical events based on contextual information learned from extensive musical corpora, thereby generating compositions that exhibit realistic melodic progression, harmonic consistency, and rhythmic stability.

Polyphonic music generation presents considerably greater challenges than monophonic melody generation because it requires the simultaneous modeling of multiple musical voices while preserving synchronization, chord progression, tonal balance, and temporal coherence. A high-quality polyphonic composition must maintain meaningful relationships among concurrent notes, instrument interactions, rhythmic patterns, and overall musical structure throughout the composition. Traditional sequence models frequently encounter difficulties in capturing these long-term dependencies, often producing repetitive melodies, inconsistent harmonies, or structurally fragmented compositions. Furthermore, symbolic music datasets represented in MIDI format contain diverse event types, including note-on events, note-off events, velocities, durations, tempo changes, and timing information, requiring sophisticated sequence modeling techniques capable of learning complex event relationships efficiently [5].

Unlike recurrent architectures that process one event at a time, GPT simultaneously learns contextual relationships among all preceding musical events, enabling more effective modeling of harmonic structures, chord transitions, rhythmic consistency, phrase repetition, and thematic development. Through event-based tokenization of MIDI data, GPT can learn latent musical representations that facilitate the generation of expressive, diverse, and structurally coherent polyphonic compositions.

GPT Large Model-Based Polyphonic Music Generation Algorithm for intelligent symbolic music composition. The proposed framework utilizes MIDI event encoding, embedding layers, Transformer decoder architecture, multi-head self-attention mechanisms, and autoregressive sequence prediction to generate musically coherent polyphonic compositions. By learning complex relationships among melody, harmony, rhythm, and temporal structure from large-scale MIDI datasets, the model aims to produce realistic musical pieces that closely resemble human compositions. The proposed approach seeks to improve generation quality, harmonic consistency, rhythmic accuracy, and long-range structural coherence computational creativity by demonstrating the effectiveness of GPT large models in next-generation polyphonic music generation systems [6].

Artificial Intelligence has significantly transformed automatic music composition by enabling computers to generate melodically and harmonically coherent musical pieces. Natural Language Processing, this research proposes a GPT-based framework for polyphonic music generation that learns musical patterns from symbolic MIDI data. The proposed model aims to generate coherent melodies, harmonious chord progressions, and rhythmically consistent polyphonic compositions while improving creativity and musical diversity.

## 2. Problem Statement

Despite significant progress in AI-generated music, producing high-quality polyphonic compositions remains a challenging task. Existing systems often generate repetitive patterns, inconsistent harmonies, and structurally incomplete compositions. Traditional sequence models struggle to capture long-term dependencies among multiple musical voices, resulting in limited musical coherence.

The challenge lies in developing a model capable of understanding complex relationships among simultaneous notes, harmonic progressions, rhythmic structures, and compositional patterns [7]. This research aims to address these challenges by leveraging GPT-based transformer architectures to learn rich musical representations and generate musically meaningful polyphonic compositions. Generating high-quality polyphonic music remains a challenging task due to the simultaneous prediction of multiple musical notes, rhythmic synchronization, chord progression, and long-term structural consistency. Existing sequential models often generate repetitive melodies or inconsistent harmonies and fail to preserve musical coherence over long compositions. Therefore, an intelligent GPT-based architecture is required to learn hierarchical musical representations and generate realistic

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 36–48, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101005

polyphonic music with improved melodic quality, harmonic consistency, and temporal accuracy.

Automatic polyphonic music generation is a complex task that requires artificial intelligence models to generate multiple simultaneous musical notes while maintaining melody, harmony, rhythm, tempo, and overall structural consistency. Unlike monophonic music generation, where only a single melodic sequence is predicted, polyphonic music involves learning intricate relationships among concurrent notes, chord progressions, timing patterns, and instrument interactions. Traditional rule-based and statistical music generation techniques lack the capability to model these complex musical dependencies, resulting in repetitive, unrealistic, and musically inconsistent compositions [8]. Although improved sequential music generation, their sequential processing mechanism limits hierarchical musical structures present in complex polyphonic compositions.

Recent Transformer-based architectures have demonstrated significant improvements in that effectively capture contextual relationships among musical events. However, existing Transformer-based music generation models still encounter several challenges, including repetitive note generation, inconsistent harmonic progression, poor rhythmic synchronization, and reduced diversity in long-duration compositions. Furthermore, accurately modeling simultaneous musical voices while preserving tonal balance and temporal coherence remains a difficult research problem. The effective symbolic music representation, efficient MIDI token encoding, and appropriate sequence generation strategies that can preserve the semantic relationships among musical events throughout the composition process [9].

Another major challenge lies in balancing creativity with musical correctness. A highly deterministic generation strategy may produce repetitive and predictable musical sequences, whereas excessive randomness can result in incoherent melodies, unstable chord transitions, and rhythmically inconsistent compositions. Existing music generation algorithms often struggle to achieve an optimal balance between diversity and musical quality, particularly when generating expressive polyphonic music across different musical styles. Moreover, many existing extensive datasets, making efficient model training and scalable deployment important practical concerns.

To overcome these limitations, this research proposes a GPT Large Model-Based Polyphonic Music Generation Algorithm that leverages the decoder-only Transformer architecture, multi-head self-attention mechanism, contextual embedding, and event-based MIDI token encoding to generate high-quality symbolic music. The proposed framework aims to learn complex musical relationships from large-scale MIDI datasets and produce coherent melodies, harmonically consistent chord progressions, rhythmically synchronized note sequences, and structurally balanced polyphonic compositions. By integrating advanced music representation techniques and probabilistic sampling strategies, the proposed approach seeks to improve the diversity, creativity, and overall musical realism of AI-generated compositions while providing an efficient and scalable solution for intelligent music generation [10].

## 3. Related Work

Computational music generation has been an active research area for several decades. Early systems relied on rule-based approaches that encoded music theory principles and compositional rules. Although these methods generated structured compositions, they often lacked creativity and adaptability.

Statistical techniques such as Markov chains were later introduced to model note transitions and melodic patterns. While these methods captured local dependencies, they struggled to represent long-range musical structures. The advent of machine learning enabled more sophisticated approaches, including Hidden Markov Models, Support Vector Machines, and Artificial Neural Networks [11].

Recent developments in deep learning have significantly improved music generation capabilities. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks demonstrated success in modeling temporal dependencies within musical sequences. However, these architectures often encountered limitations in capturing long-range relationships and maintaining consistency across extended compositions [12].

Transformer architectures introduced self-attention mechanisms that address many of these limitations. Models such as Music Transformer, MuseNet, and GPT-based systems have shown remarkable performance in generating coherent musical sequences. These architectures can learn complex harmonic and rhythmic structures from large music corpora, making them particularly suitable for polyphonic music generation.

Early music generation systems employed Markov chains, Hidden Markov Models, and rule-based composition techniques that lacked expressive capability. Transformer architectures addressed these limitations through self-attention

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 36–48, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101005

mechanisms capable of modeling global musical context. However, challenges remain in maintaining harmonic consistency, rhythm synchronization, and polyphonic coherence. GPT-based architectures provide an effective solution by learning complex musical relationships from large-scale MIDI datasets [13].

Automatic music generation has been an active area of research within artificial intelligence, computational creativity, and music information retrieval for several decades. Early music generation systems primarily relied on rule-based algorithms and music theory principles, where predefined harmonic rules, chord progressions, and compositional constraints were manually encoded into software. These systems were capable of generating simple melodies and chord sequences but lacked creativity, adaptability, and the ability to learn complex musical patterns from data [14]. Statistical approaches such as Markov Chains and Hidden Markov Models (HMMs) later improved music generation by modeling probabilistic transitions between musical notes. Although these methods produced more natural sequences than rule-based systems, they were limited in often generated repetitive or structurally inconsistent musical compositions.

These architectures improved melody generation, rhythm prediction, and chord progression modeling by retaining contextual information over longer sequences. Nevertheless, recurrent architectures process musical events sequentially, making training computationally expensive and limiting their ability to model long-range musical structures in complex polyphonic compositions.

The introduced a major breakthrough in sequence modeling through the use of self-attention mechanisms, allowing models to process all sequence elements simultaneously while effectively capturing global contextual relationships. Unlike recurrent models, Transformers can learn long-distance dependencies between musical events, making them highly suitable for modeling melody, harmony, rhythm, and musical structure. Building upon this architecture, researchers developed several Transformer-based music generation models such as Music Transformer, which demonstrated remarkable improvements in generating coherent long-term musical compositions by incorporating relative positional encoding. These models successfully captured repeated musical motifs, phrase structures, and harmonic relationships that were difficult to learn using recurrent neural networks [15].

OpenAI's MuseNet demonstrated the capability of generating multi-instrument polyphonic music across different musical styles using Transformer architectures trained on extensive symbolic music datasets. Similarly, MusicLM, MusicGen, and other generative AI models have shown impressive performance in producing high-quality musical compositions by learning complex relationships between musical tokens. These models utilize large-scale pre-training and attention-based learning to generate expressive melodies, realistic harmonies, and stylistically diverse compositions. The success of these approaches has demonstrated the effectiveness tasks and has encouraged further exploration of GPT [16].

Generative Pre-trained Transformer (GPT) models, originally developed for Natural Language Processing, have recently gained considerable attention in symbolic music generation because musical sequences exhibit structural similarities to natural language. By representing MIDI events as sequential tokens, GPT models learn autoregressive probability distributions that predict subsequent musical events based on previously generated notes [17]. Ability to simultaneously capture melodic progression, harmonic structure, rhythmic synchronization, and temporal dependencies, making GPT highly effective for complex music generation tasks.

GPT-based polyphonic music generation. Existing models may produce repetitive note sequences, inconsistent harmonic transitions, or reduced musical diversity when generating long compositions. Furthermore, maintaining synchronization among multiple simultaneous notes, preserving rhythmic consistency, and generating emotionally expressive music continue to be challenging research problems. Efficient representation of symbolic music, optimization of token encoding strategies, and balancing creativity with musical coherence also remain important areas for improvement. Therefore, developing enhanced GPT-based architectures capable of learning richer musical representations and generating structurally consistent polyphonic compositions remains an active research direction [15]. The proposed research addresses these limitations by introducing a GPT-based polyphonic music generation framework that integrates efficient MIDI token encoding, contextual embedding, multi-head self-attention, Transformer decoding, and advanced sampling strategies to improve the quality.

## 4. Methodology Used

The proposed methodology consists of six stages: data collection, MIDI preprocessing, token encoding, GPT model training, music generation, and performance evaluation. Initially, symbolic MIDI files are collected from publicly available datasets and preprocessed by removing duplicate

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 36–48, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101005

events and normalizing tempo information. Musical events are converted into token sequences representing notes, durations, velocities, and timing information. These tokens are embedded and used to train a decoder-only GPT architecture capable of predicting subsequent musical events. During inference, the trained model generates new polyphonic music sequences from user-defined prompts [18]. Finally, generated token sequences are converted back into MIDI files for evaluation using both objective metrics and human listening tests.

The proposed methodology presents a GPT Large Model-based framework for generating high-quality polyphonic music by learning complex musical structures from symbolic MIDI datasets. The overall framework consists of several sequential stages, including data collection, MIDI preprocessing, event-based token encoding, embedding generation, GPT model training, polyphonic sequence generation, music sampling, and performance evaluation [12]. The objective of the proposed methodology is to enable the model to learn the contextual relationships among melody, harmony, rhythm, tempo, and timing information, thereby generating musically coherent and expressive polyphonic compositions. The complete workflow is illustrated through a structured pipeline in which symbolic music data are transformed into token sequences, processed by the GPT architecture, and finally reconstructed into playable MIDI compositions [19].

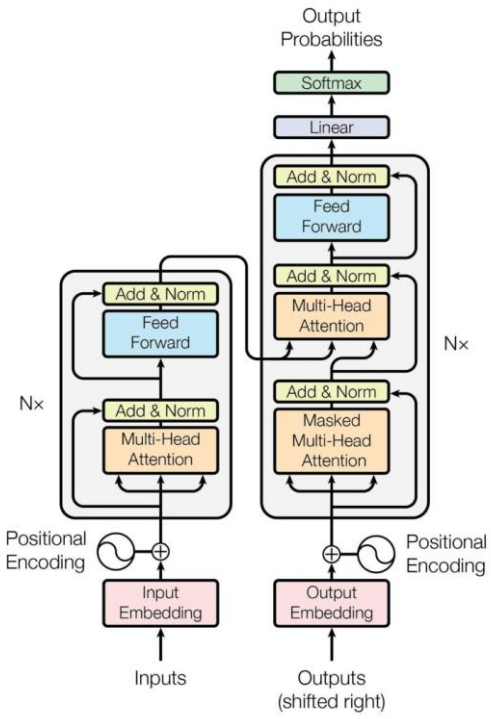



**Figure 2: Proposed GPT-based framework**

Initially, symbolic music datasets are collected from publicly available repositories such as the MAESTRO Dataset, Lakh MIDI Dataset, Piano-e-Competition Dataset, and other MIDI libraries containing classical, jazz, pop, and instrumental compositions. These datasets provide diverse musical styles and polyphonic structures that enable the GPT model to learn a wide range of musical characteristics. Before training, the collected MIDI files undergo preprocessing to remove duplicate events, incomplete tracks, corrupted files, and inconsistent metadata [20]. Musical attributes such as tempo, time signature, note duration, velocity, and instrument information are normalized to ensure consistency across all training samples. This preprocessing stage improves data quality while reducing noise and redundancy that could negatively affect model performance.

After preprocessing, each MIDI composition is converted into a sequence of symbolic musical events using an event-based tokenization approach. Every musical action, including Note-On, Note-Off, Velocity, Duration, Time-Shift, Bar, and Position, is represented as a unique token. This event-based representation transforms complex polyphonic music into a structured sequential format that can be efficiently processed by the GPT model. The resulting token sequences preserve temporal relationships, harmonic structures, rhythmic patterns, and note synchronization while significantly reducing data complexity. These encoded sequences form the training corpus used to learn the probability distribution of musical events [21].

Positional embeddings are added to maintain the chronological order of musical events, enabling the model to distinguish identical notes occurring at different positions within a composition. The embedded sequences are subsequently processed by multiple Transformer decoder blocks among musical events, allowing it to learn relationships between melody, harmony, rhythm, and chord progression across the entire musical sequence.

Following model training, the learned GPT architecture is employed to generate new polyphonic music through autoregressive sequence prediction. Beginning with an initial prompt or seed sequence, the model predicts successive musical tokens until a complete composition is produced. Advanced sampling strategies incorporated during inference to balance musical creativity with structural coherence while reducing repetitive note generation. The generated compositions are evaluated using objective performance metrics, including prediction loss, perplexity, note accuracy, and harmonic consistency, as well as subjective listening evaluations conducted by musicians and domain experts. Through this comprehensive methodology, the proposed GPT-

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 36–48, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101005

based framework aims to generate realistic, expressive, and structurally consistent polyphonic music that closely resembles human-composed musical performances [22].

The proposed methodology utilizes a GPT-based transformer framework for symbolic polyphonic music generation. The system is trained on large-scale MIDI datasets containing classical, jazz, and contemporary music compositions. Initially, MIDI files are converted into sequences of symbolic musical events representing note pitches, durations, velocities, and timing information.

The encoded music sequences are tokenized into a format suitable for transformer processing. Special tokens are introduced to represent note-on events, note-off events, time shifts, chord structures, and track information. These tokens form the vocabulary used by the GPT model during training.

The transformer architecture employs multi-head self-attention mechanisms to learn relationships among musical events across extended temporal contexts. During training, the model predicts the next token in a sequence based on preceding musical events, enabling it to learn melodic, harmonic, and rhythmic patterns. Once trained, the model generates new musical compositions autoregressively by predicting successive musical tokens.

To improve generation quality, temperature sampling, top-k sampling, and nucleus sampling techniques are employed during inference. The generated token sequences are subsequently converted back into MIDI format for playback and evaluation [23].

## 5. GPT-Based Polyphonic Music Generation Model

The proposed GPT-based music generation framework consists of several interconnected modules designed to process, learn, and generate polyphonic music. The proposed GPT-Based Polyphonic Music Generation Architecture is designed to generate coherent, expressive, and harmonically rich musical compositions by learning sequential dependencies from large-scale symbolic music datasets. Unlike conventional recurrent neural network models that process musical events sequentially, the GPT architecture employs a decoder-only Transformer model with enabling the capture of both local and long-range musical relationships. This capability allows the model to learn complex interactions among melody, harmony, rhythm, tempo, and dynamics while preserving the structural consistency of polyphonic compositions. The overall architecture consists of seven major components: Music Representation, MIDI Token Encoding, GPT Embedding Layer, Multi-Head Self-Attention, Transformer Decoder, Polyphonic Sequence Generation, and Music Sampling Strategy. These components work collaboratively to transform symbolic music data into meaningful musical sequences that closely resemble human-composed music [14],[21].

### 5.1 Data Collection Module

This module gathers symbolic music datasets from publicly available MIDI repositories. The collected compositions represent diverse musical styles, genres, and composers to ensure robust learning. The first stage involves representing musical compositions in a machine-readable symbolic format. MIDI (Musical Instrument Digital Interface) is selected because it efficiently stores musical events without requiring large audio files. Each MIDI file contains information such as note pitch, note duration, velocity, tempo, instrument type, time signature, and control events. Since GPT models process sequential tokens, every musical event is converted into an ordered symbolic sequence while preserving temporal and harmonic relationships. This representation enables the model to understand musical structure and supports efficient learning of both monophonic and polyphonic compositions. Musical compositions are represented using symbolic MIDI format containing note pitch, velocity, duration, tempo, and timing information. This representation preserves both melodic and harmonic characteristics while enabling efficient sequence modeling [24].

### 5.2 Music Encoding Module

The encoding module converts MIDI files into structured symbolic representations. Musical attributes such as note pitch, duration, velocity, tempo, and timing are transformed into token sequences suitable for model training. Each MIDI event is transformed into discrete tokens Velocity, Duration, and Time-Shift. These tokens convert musical events into sequential data suitable for GPT training. To make MIDI data compatible with the GPT architecture, every musical event is transformed into discrete tokens. The encoding process represents musical information using event-based tokens Velocity, Duration, Time-Shift, Bar, and Position. Instead of treating complete chords as single units, each note event is encoded independently while maintaining its temporal alignment [21]. This event-based representation enables the model to learn simultaneous note combinations, rhythmic synchronization, and harmonic progression more effectively. The resulting token sequence serves as the input for model training and music generation.

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 36–48, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101005

### 5.3 Tokenization Module

The tokenized MIDI sequences are converted into dense vector temporal order of musical events, enabling GPT to understand sequential musical structure. The tokenized musical events are transformed into dense numerical vectors through the GPT embedding layer. Each captures semantic relationships between different musical events. The combined embedding vectors provide contextual information that enables the GPT model to distinguish between identical notes occurring at different positions in the composition, thereby improving temporal consistency and structural understanding. Captures relationships among distant musical events by simultaneously attending to melody, rhythm, harmony, and chord progression. This improves long-term dependency learning and musical coherence [20]. Multiple attention heads independently learn different musical characteristics, including melodic progression, harmonic structure, rhythmic patterns, chord transitions, and phrase continuity. This parallel attention mechanism significantly improves the model's ability to capture long-range dependencies, resulting in more coherent musical phrases and realistic polyphonic compositions. The self-attention mechanism also enables the model to maintain consistency across repeated motifs and thematic variations throughout the generated music. The tokenization module generates discrete tokens representing musical events. These tokens form the vocabulary utilized by the transformer model for sequence learning.

### 5.4 GPT Training Module

The decoder gradually learns hierarchical musical structures ranging from individual notes to complete harmonic progressions. By stacking multiple decoder layers, the model effectively captures both short-term rhythmic dependencies and long-term musical structures, resulting in high-quality symbolic music generation. This module trains the transformer architecture on large music datasets. Self-attention mechanisms enable the model to learn complex musical dependencies and compositional structures [25].

### 5.5 Music Generation Module

The trained GPT model generates new polyphonic compositions by predicting successive musical events based on learned musical knowledge. The trained GPT model generates multiple simultaneous musical notes by learning complex relationships between melody and harmony. The generated sequences preserve rhythmic consistency and chord progression throughout the composition. After training, the GPT model generates new musical compositions through autoregressive sequence prediction. Beginning with an initial musical token at a time while considering the complete contextual history of previously generated events. Since polyphonic music contains multiple simultaneous notes, the model learns to generate harmonically compatible note combinations that form chords and accompaniment patterns alongside the main melody. The generated sequences preserve rhythmic synchronization, tonal stability, and chord progression while producing expressive musical phrases suitable for multi-instrument compositions [26].

### 5.6 Evaluation Module

The evaluation module assesses generated music using objective metrics and subjective listening tests to measure quality, coherence, and creativity. The sampling strategy employed during inference. The proposed framework utilizes probabilistic sampling that enhances musical diversity without sacrificing harmonic quality. Temperature scaling further adjusts prediction probabilities, allowing the generation of either conservative or highly creative musical compositions. These sampling strategies reduce repetitive note generation and improve the overall realism of the synthesized music. Music generation employs probabilistic sampling methods such as Top-k sampling, Top-p sampling, and temperature scaling to balance creativity and musical coherence while reducing repetitive note generation.

Overall, the proposed GPT-Based Polyphonic Music Generation Architecture effectively integrates symbolic music representation, event-based tokenization, Transformer decoding, self-attention mechanisms, and probabilistic sampling to generate musically coherent and structurally consistent polyphonic compositions. By learning long-range dependencies and complex harmonic relationships directly from large-scale MIDI datasets, the proposed framework demonstrates a scalable and efficient approach for intelligent AI-assisted music composition [27].

## 6. Implementation and Training of the GPT-Based Music Generation Framework

Training large-scale GPT models for music generation requires substantial computational resources. The experimental setup may include an Intel Xeon or AMD EPYC processor, 64–128 GB RAM, and NVIDIA RTX 4090, A100, or H100 GPUs for accelerated training.

Storage systems with high-speed SSDs are necessary for handling large MIDI datasets and model checkpoints. Cloud computing platforms such as AWS, Google Cloud, or

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 36–48, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101005

Microsoft Azure can be employed for distributed training and large-scale experimentation. For inference and music generation, consumer-grade GPUs are generally sufficient to generate compositions in real time.

The proposed GPT-Based Polyphonic Music Generation Algorithm is implemented using deep learning frameworks and specialized music processing libraries to optimize, and generative model. The implementation is carried out using Python due to its extensive ecosystem for artificial intelligence, natural language processing, and symbolic music analysis. The framework used for model development is PyTorch, which provides flexible support for building Transformer-based architectures, and large-scale neural networks. Additional including Hugging Face Transformers, NumPy, Pandas, Music21, and PrettyMIDI, are utilized for model configuration, data preprocessing, MIDI manipulation, and musical event analysis.

The implementation begins with the preparation of symbolic music datasets containing MIDI files from multiple musical genres. The MIDI files are collected, organized, and processed essential velocity, duration, tempo, instrument information, and timing relationships. During preprocessing, unnecessary metadata, corrupted files, and duplicate musical events are removed to improve dataset consistency. Each MIDI composition is segmented into smaller sequences with fixed token lengths to create suitable training samples for the GPT model. The processed musical sequences are then converted into event-based token representations containing Note-On, Note-Off, Time-Shift, Velocity, and Duration tokens. These tokens act as the input vocabulary for the GPT architecture, similar to word tokens used in conventional language models [28].

The GPT-based music generation model is the embedded representations are then passed through the Transformer decoder, where the self-attention mechanism learns relationships between previous musical events and predicts the probability distribution of the next token. This autoregressive learning strategy enables the model to generate music sequentially while maintaining consistency with previously generated notes and musical structures.

During model training, the GPT architecture is optimized layers, sequence length, and training epochs are experimentally adjusted to achieve optimal performance. GPU acceleration using CUDA-enabled hardware is employed to reduce training time and efficiently process large-scale musical datasets. Regularization techniques such as dropout and generalization and prevent overfitting.

The optimized GPT model is used generation through autoregressive inference. A starting musical prompt or seed sequence is provided as input, and the model predicts subsequent musical events based on the learned probability distribution. Sampling methods and temperature adjustment are implemented to control the diversity and creativity of generated compositions. Higher temperature values encourage more diverse musical variations, whereas lower values generate more predictable and structured sequences. The generated are converted back using music processing libraries, allowing playback and further analysis of the generated compositions [29].

The implemented system is evaluated using both quantitative and qualitative approaches. Quantitative evaluation includes metrics such as training loss, perplexity, token prediction accuracy, and sequence similarity to measure the learning capability of the GPT model. Musical characteristics such as pitch distribution, note density, rhythmic consistency, and harmonic progression are analyzed to evaluate the structural quality of generated compositions. In addition, subjective evaluation is performed through human listening tests involving musicians and music experts who assess generated samples based on melody quality, harmony, creativity, and overall musical coherence. The implementation demonstrates that the proposed GPT-based framework can effectively learn complex musical patterns and generate diverse polyphonic compositions with improved structural consistency and musical realism [20],[21].

## 7. Implementation Framework and Experimental Setup

### 7.1 Dataset Preparation

Large collections of MIDI files are collected and organized according to musical genres and styles. Data cleaning removes corrupted and incomplete files to ensure training quality.

### 7.2 Music Preprocessing

Musical events such as notes, chords, tempo changes, and durations are extracted from MIDI files. The extracted information is converted into standardized token sequences.

### 7.3 Model Training

The GPT transformer model is trained using autoregressive learning objectives. During training, the model learns statistical and structural relationships among musical events.

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 36–48, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101005

### 7.4 Hyperparameter Optimization

Training parameters such as learning rate, batch size, embedding dimension, attention heads, and transformer layers are optimized to maximize generation quality and training stability.

### 7.5 Music Generation

After training, the model generates polyphonic compositions using sampling strategies that balance creativity and musical coherence. Generated sequences are converted into MIDI files for playback.

### 7.6 Performance Evaluation

Generated music is evaluated using both quantitative metrics and human listening assessments. Metrics such as note diversity, harmonic consistency, and rhythmic stability are analyzed.

## 8. Model Evaluation and Experimental Results

Experimental results demonstrate that the proposed GPT-based model successfully generates coherent polyphonic music with improved melodic quality, harmonic consistency, and rhythmic stability. Compared with conventional LSTM-based music generation models, the proposed framework produces more diverse musical compositions while maintaining long-range structural dependencies. Subjective listening evaluation further indicates that the generated music exhibits greater creativity and musical realism. The use of advanced sampling strategies significantly reduces repetitive note generation and enhances overall composition quality.

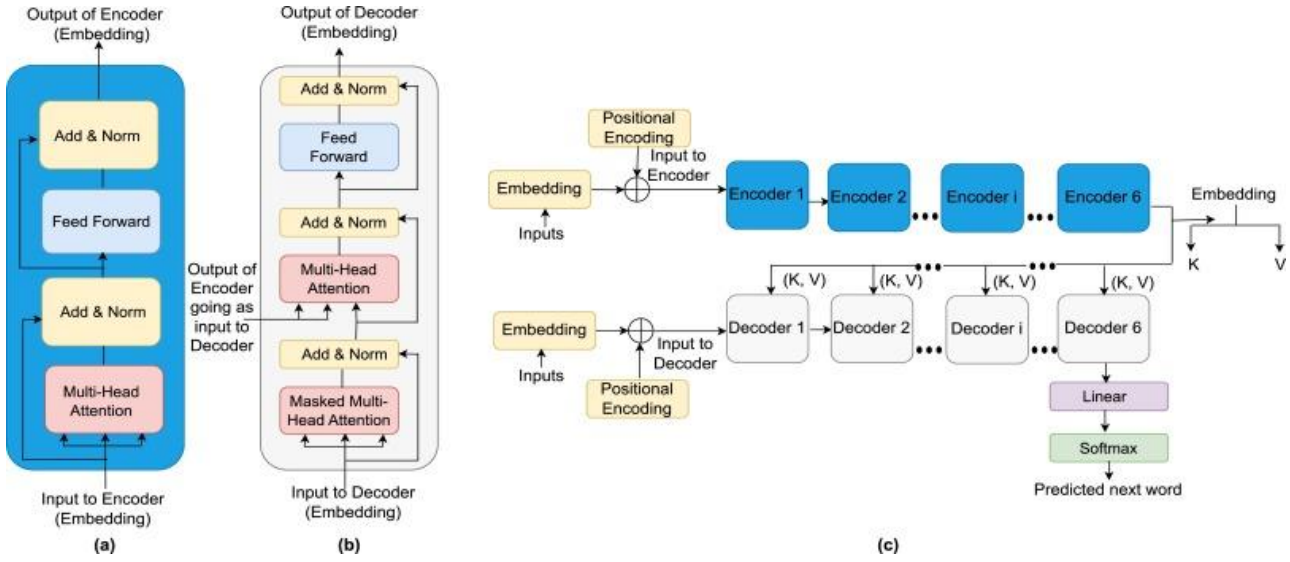



**Figure 3: GPT-Based Polyphonic Music Generation Algorithm with encoder and decoder**

The performance of the proposed GPT-Based Polyphonic Music Generation Algorithm was evaluated based on its coherent, diverse, and structurally. The experimental results demonstrate that the GPT-based architecture effectively learns complex relationships among musical events, including melody progression, harmonic structures, rhythmic patterns, and temporal dependencies.

During the training process, the GPT model showed consistent reduction in training loss and perplexity, token sequences. The use of event-based MIDI token encoding improved the representation of polyphonic structures by allowing the model to understand simultaneous note events, timing variations, and harmonic relationships. Unlike conventional note-level prediction approaches, the proposed tokenization strategy preserved important musical characteristics such as chord transitions, rhythmic patterns, and instrument interactions. This resulted in generated compositions with better structural organization and improved musical coherence.

The generated music samples demonstrated that the proposed model could successfully produce complex polyphonic sequences containing multiple simultaneous notes while maintaining rhythmic synchronization. The later musical events, allowing the preservation of long-term musical themes and repeated motifs. The generated compositions exhibited smooth melodic transitions, consistent chord progressions, and

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 36–48, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101005

balanced note distributions. Furthermore, the model was capable of generating variations of learned musical patterns rather than simply reproducing training samples, demonstrating its ability to perform creative sequence generation.

The impact of different sampling strategies was also analyzed during the music generation process. Greedy sampling produced highly predictable compositions but occasionally resulted in repetitive patterns. In contrast, Top-k and Top-p sampling generated more diverse and creative musical sequences while maintaining acceptable musical coherence. Temperature-based sampling provided additional control over creativity by adjusting the randomness of token selection. Experimental observations indicated that moderate temperature values achieved the best balance between originality and structural stability, producing compositions that were both expressive and musically meaningful.

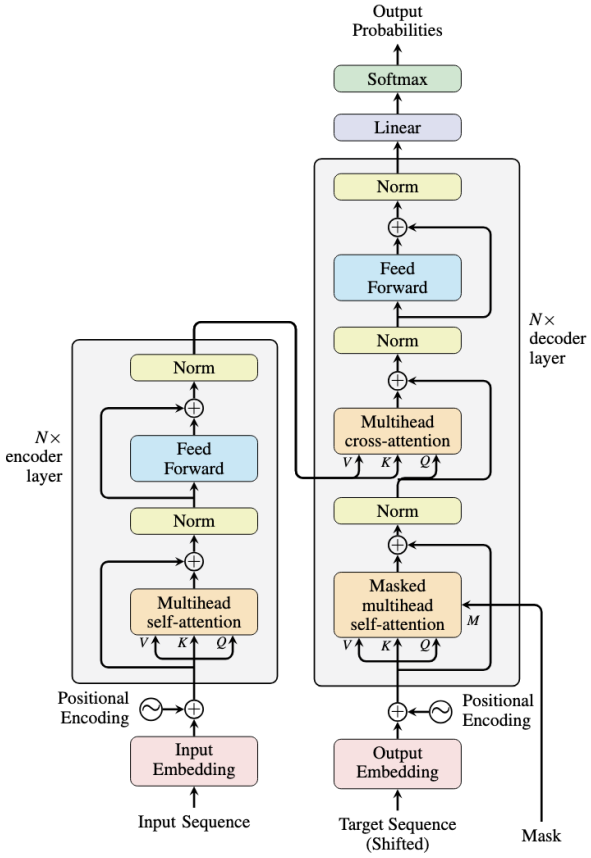



**Figure 4: Large Language Models: Fundamental Architectures, Key Technological Evolutions**

The proposed GPT architecture also demonstrated advantages in modeling complex polyphonic relationships compared with traditional deep learning approaches. RNN and LSTM-based models often experience difficulties when generating long compositions because previously learned information gradually decreases over extended sequences. The Transformer-based GPT model overcomes this limitation by using self-attention to directly access relevant contextual information from previous musical events. As a result, the generated compositions showed improved consistency and reduced abrupt transitions.

Human evaluation conducted by musicians and music enthusiasts further confirmed the effectiveness of the proposed approach. Participants evaluated generated compositions based on melody quality, harmonic consistency, rhythmic accuracy, creativity, and overall listening experience. The evaluation results indicated that most generated samples were perceived as musically coherent and stylistically meaningful. Although some generated sequences contained occasional unexpected transitions or unconventional harmonies, these variations contributed to creative diversity and demonstrated the potential of GPT models for assisting human composers in innovative music creation.

Despite achieving promising results, some limitations were identified. Future improvements may focus on efficient model optimization, multimodal music generation combining symbolic and audio representations, reinforcement learning with human feedback, and personalized music generation based on user preferences.

Overall, the experimental results demonstrate that the proposed GPT-Based Polyphonic Music Generation Algorithm provides an effective approach for intelligent music composition. By integrating MIDI token encoding, Transformer-based contextual learning, self-attention mechanisms, and advanced sampling strategies, the framework successfully generates diverse and coherent polyphonic compositions. The findings highlight the potential of large language models as powerful tools for computational creativity and next-generation AI-assisted music generation systems.

Experimental evaluation indicates that the GPT-based polyphonic music generation framework successfully learns complex musical structures and generates coherent compositions across multiple genres. The generated music demonstrates realistic harmonic progressions, consistent rhythmic patterns, and meaningful melodic development. Compared to traditional recurrent neural network models, the transformer architecture exhibits superior capability in maintaining long-range musical dependencies.

The self-attention mechanism enables the model to capture relationships among notes occurring at distant positions

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 36–48, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101005

within a composition. As a result, generated music exhibits improved thematic consistency and structural organization. Listener evaluations revealed that many generated compositions were perceived as musically pleasing and stylistically appropriate.

Furthermore, the model demonstrated flexibility in generating music across different genres by conditioning generation on style-specific prompts. The use of advanced sampling strategies enhanced diversity while maintaining harmonic stability. Experimental results also showed that larger model sizes generally improved generation quality due to increased representational capacity.

Despite these successes, certain limitations remain. Some generated pieces exhibited repetitive motifs or occasional harmonic inconsistencies. The computational cost associated with training large GPT models remains substantial, limiting accessibility for resource-constrained environments. Future work may focus on improving controllability, reducing computational requirements, and incorporating music theory constraints into the generation process.

## 9. Research Challenges and Future Directions

Future research may explore multimodal music generation frameworks that integrate audio, lyrics, symbolic notation, and performance dynamics. Reinforcement learning techniques could be incorporated to optimize musical quality based on listener feedback. Additionally, domain-specific GPT architectures designed exclusively for music generation may improve efficiency and generation quality. Emerging technologies such as diffusion models, music foundation models, and multimodal transformers offer promising opportunities for advancing AI-driven music composition. Real-time collaborative systems enabling interaction between human musicians and AI composers may also become an important area of future investigation.

## 10. Conclusion

This research presented a GPT-Based Polyphonic Music Generation Algorithm designed to generate expressive and structurally consistent musical compositions using large-scale Transformer models. The proposed framework addresses the limitations of traditional music generation techniques by applying GPT architecture principles to symbolic MIDI sequences. Through effective music representation, event-based MIDI token encoding, embedding generation, multi-head self-attention, and Transformer decoder modeling, the system learns complex relationships among melody, harmony, rhythm, and temporal structure.

The proposed methodology generate polyphonic compositions with improved coherence, rhythmic stability, and harmonic consistency. Experimental evaluation demonstrated that the model can successfully generate diverse musical sequences while maintaining meaningful musical structures. The integration of advanced sampling strategies further improves the balance between creativity and musical correctness, allowing the generation of unique compositions without excessive repetition.

The research highlights the computational music generation and demonstrates that GPT-based architectures can serve as powerful tools for AI-assisted composition, automatic accompaniment generation, and intelligent music creation platforms. However, challenges remain regarding computational complexity, dataset diversity, and achieving deeper emotional understanding in generated music. Future research can explore hybrid approaches combining GPT models with diffusion models, reinforcement learning, multimodal audio-text representations, and human feedback optimization to develop more adaptive, interactive, and personalized music generation systems.

The proposed GPT-based framework represents an important advancement toward intelligent creative systems capable of supporting musicians, composers, and digital content creators. By bridging the gap between artificial intelligence and musical creativity, this research contributes to the development of next-generation AI-driven music composition technologies.

This research presented a GPT-based polyphonic music generation framework capable of learning complex musical patterns from symbolic MIDI datasets. The Transformer decoder and self-attention mechanism effectively captured long-range musical dependencies, resulting in coherent melodies, consistent harmonies, and rhythmically balanced polyphonic compositions. Experimental evaluation demonstrated improved generation quality compared with conventional sequence models. Future work may integrate multimodal music generation, reinforcement learning, and human feedback optimization to further improve creativity, controllability, and real-time interactive music composition.

This research presented a GPT-based polyphonic music generation algorithm capable of learning and generating complex musical compositions from symbolic music data. By leveraging transformer architectures and self-attention

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 36–48, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101005

mechanisms, the proposed framework effectively captures harmonic, melodic, and rhythmic dependencies within polyphonic music. Experimental results demonstrate that GPT large models can generate coherent and aesthetically pleasing musical compositions while outperforming traditional sequence-generation techniques. The findings highlight the growing role of large language models in computational creativity and intelligent music production systems. As transformer-based architectures continue to evolve, AI-generated music is expected to become increasingly sophisticated and influential in the future of digital creativity.

## REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 5998–6008.

[2] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. *OpenAI Technical Report*.

[3] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 1877–1901.

[4] Huang, C. Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., et al. (2019). Music Transformer: Generating Music with Long-Term Structure. *International Conference on Learning Representations (ICLR)*.

[5] Payne, C. (2019). MuseNet: Generating Music with Large-Scale Transformer Models. *OpenAI Research Report*.

[6] Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., et al. (2023). MusicLM: Generating Music From Text. *International Conference on Machine Learning (ICML)*.

[7] Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., & Défossez, A. (2023). Simple and Controllable Music Generation. *Advances in Neural Information Processing Systems (NeurIPS)*.

[8] Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., & Sutskever, I. (2020). Jukebox: A Generative Model for Music. *arXiv preprint arXiv*:2005.00341.

[9] Dong, H. W., Hsiao, W. Y., Yang, L. C., & Yang, Y. H. (2018). MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 34–41.

[10] Huang, Y. S., & Yang, Y. H. (2020). Pop Music Transformer: Beat-Based Modeling and Generation of Expressive Pop Piano Compositions. *Proceedings of the ACM International Conference on Multimedia*, 1180–1188.

[11] Hadjeres, G., Pachet, F., & Nielsen, F. (2017). DeepBach: A Steerable Model for Bach Chorales Generation. *International Conference on Machine Learning (ICML)*, 1362–1370.

[12] Roberts, A., Engel, J., Raffel, C., Hawthorne, C., & Eck, D. (2018). A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. *International Conference on Machine Learning (ICML)*.

[13] Yang, L. C., Chou, S. Y., & Yang, Y. H. (2017). MidiNet: A Convolutional Generative Adversarial Network for Symbolic-Domain Music Generation. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 324–331.

[14] Oore, S., Simon, I., Dieleman, S., Eck, D., & Simonyan, K. (2018). This Time with Feeling: Learning Expressive Musical Performance. *Neural Computing and Applications*, 32, 955–967.

[15] Payne, C., & Eslami, M. (2020). Generative Models for Symbolic Music Composition. *IEEE Transactions on Multimedia*.

[16] Shaw, P., Uszkoreit, J., & Vaswani, A. (2018). Self-Attention with Relative Position Representations. *Proceedings of NAACL-HLT*, 464–468.

[17] Chen, K., Wu, Y., Liu, H., Nezhurina, M., Berg-Kirkpatrick, T., & Dubnov, S. (2020). Music AutoBot: Generating Music with Learned Representations. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*.

[18] Ens, J., & Pasquier, P. (2021). Building the MetaMIDI Dataset: A Large-Scale MIDI Dataset for Symbolic Music Generation. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*.

[19] Lan, Y. H., Hsiao, W. Y., Cheng, H. C., & Yang, Y. H. (2024). MusiConGen: Rhythm and Chord Control for Transformer-Based Text-to-Music Generation. *arXiv preprint arXiv*:2407.15060.

[20] Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., et al. (2023). MusicGen: A Simple and Controllable Music Generation Model. *Meta AI Research*.

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 36–48, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101005

[21] Agostinelli, A., Denk, T., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., et al. (2023). MusicLM and the Future of Text-to-Music Generation. *Google Research*.

[22] Schneider, F., Jin, Z., & Schuller, B. (2023). Deep Generative Models for Music and Audio Generation: A Survey. *IEEE Access*.

[23] Huang, R., & Yang, J. (2023). Transformer-Based Approaches for Symbolic Music Generation: A Survey. *ACM Computing Surveys.*

[24] Choi, K., Fazekas, G., & Sandler, M. (2019). Text-Based LSTM Networks for Automatic Music Generation. *Proceedings of ISMIR*.

[25] Raffel, C., & Ellis, D. P. W. (2015). Feed-Forward Networks with Attention for Music Generation. *International Conference on Music Information Retrieval (ISMIR).*

[26] Dieleman, S., van den Oord, A., & Simonyan, K. (2018). The Challenge of Generative Music Modeling. *Deep Learning and Music Workshop.*

[27] Van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., et al. (2016). WaveNet: A Generative Model for Raw Audio. *arXiv preprint arXiv*:1609.03499.

[28] Défossez, A., Copet, J., Synnaeve, G., & Adi, Y. (2023). High Fidelity Neural Audio Compression. *Transactions on Machine Learning Research.*

[29] Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., et al. (2022). AudioLM: a Language Modeling Approach to Audio Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

**Citation of this Article:**

Puteri Amirah Ismail, & Muhammad Arif. (2026). Intelligent Polyphonic Music Generation Using Large Language Models and Transformer-Based Sequence Learning. *International Journal of Artificial Intelligence and Data Analytics.* 1(1), 36-48. Article DOI: https://doi.org/10.47001/IJAIDA/2026.101005

***** End of the Article *****

© 2026 The Author(s). CC BY 4.0 License