AI-Generated Piano Music: A Music-Theoretical Analysis and Model

James Nguyen

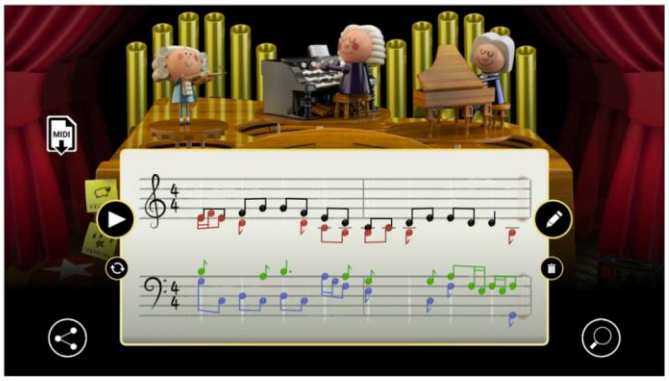
April 3, 2025

Abstract

Recent advances in artificial intelligence have enabled the generation of convincing musical compositions. This project set out to explore the music-theoretical understanding in AI-generated piano music. The initial plan was to train a neural network using Google Magenta's Performance RNN or Music Transformer on the MAESTRO dataset of piano performances and then evaluate the outputs for musical coherence. However, technical difficulties in configuring Magenta shifted the scope toward analyzing outputs from established AI music models instead. Compositions generated by OpenAI's MuseNet, Google's Bach Doodle (powered by the Coconet model), and Google's MusicLM were examined using a combination of qualitative listening and computational analysis via the music21 library. The evaluation focused on harmonic elements (including chord progressions and cadences), key consistency, melodic motifs, and overall structural coherence. Results show that today's models can display rudimentary adherence to music theory, generating plausible chord progressions and stylistic melodies though they often lack long-term structure and deep thematic development. MuseNet, for example, can convincingly blend styles and maintain a stable key for extended stretches, yet it may wander without a clear directional form [1], [2]. The Bach Doodle follows Baroque harmony rules for short chorale-style pieces but within a narrow scope. MusicLM demonstrates impressively coherent audio extending over minutes [3], although assessing its theoretical correctness demands careful listening. These findings highlight both the promise and the limitations of current AI music systems in reflecting authentic music-theoretical principles. This draft documents the mid-project pivot and presents an initial comparative analysis, serving as a foundation for a more comprehensive final report. The discussion also considers how these model outputs relate to broader questions of creativity, authorship, and the depth of AI's "understanding" of music theory.

Visual Abstract

Figure: The interface of Google's Bach Doodle (2019) provides a clear example of AI-assisted composition. In this tool, a user-submitted melody (top staff, shown in red/green) is automatically harmonized into a four-part Bach-style chorale (additional notes in blue) by the Coconet model. Although the initial plan for this project involved training a new model on piano data, the scope shifted to analyzing the outputs of such pre-existing AI systems instead. Musical quality is evaluated by examining how well the generated pieces adhere to harmonic and structural principles, including proper cadences, consistent key centers, and motivic development.



Introduction

Generative AI has rapidly progressed from creating text and images to composing music. AI-generated music presents a fascinating testbed for examining how well machine learning models capture the structure and rules of an abstract domain, such as music. Classical Western music has well-defined theoretical frameworks (tonal harmony, voice leading, musical form). The question motivating this project is: Do AI music models implicitly learn these music-theoretical principles, and to what extent do their compositions exhibit musical coherence akin to human-composed music?

Prior work has shown both promise and challenges. Early neural music systems like Magenta's Performance RNN (an LSTM for piano performances) could produce expressive short passages but often sounded like aimless "noodling" without long-term direction [4]. In other words, the model captured local phrasing and dynamics but lacked higher-level structure. Researchers identified that music contains hierarchical patterns from motifs and phrases up to entire sections [5], and that sequence models need to handle long-range dependencies to create coherent pieces. This led to the development of models such as Music Transformer, which uses self-attention with a relative positioning mechanism to better maintain musical context over longer spans [5]. Indeed, Huang et al. (2018) reported improved long-term coherence in generated piano music using the Music Transformer, compared to the RNN baseline.

Despite such progress, achieving human-like musical form remains difficult. A recent systematic review by Civit et al. (2022) notes that transformer-based models (e.g., Google's Music Transformer and OpenAI's MuseNet) generally produce more complex and structured music than earlier approaches [2]. However, even these advanced models can struggle with long-term structure: for example, a model may meander or repeat itself after a certain length, lacking the clear development or resolution a human composer might craft [2], [4]. Evaluating the quality of AI-composed music is itself an open challenge, as it involves subjective judgment and context. Past studies have used metrics from music theory (such as rule violations in Bach-style harmonies) and listener studies to assess output quality.

Project Goals: Within this backdrop, the project aimed to probe the music-theoretical competence of AI in generating piano music. The plan was initially to train a new model on a high-quality piano performance dataset and analyze its outputs for theoretical soundness (e.g., do the AI compositions contain sensible chord progressions? Do they stick to a key? Do they use cadences to punctuate phrases?). The underlying hypothesis was that if a model truly learns musical structure, its outputs should reflect the patterns and constraints described by music theory, not just surface-level continuity.

Scope Shift: Unfortunately, technical issues were encountered in setting up the training environment for Magenta’s models. Compiling and running the Performance RNN/Music Transformer in Visual Studio Code proved more challenging than anticipated, given dependency conflicts and time constraints. With a draft deadline of April 3 looming, a decision was made to pivot the project's scope. Instead of training a model from scratch, this approach leverages existing state-of-the-art AI music systems to generate music and focuses on analyzing those outputs. This change still addresses the core research questions using readily available model outputs, albeit without the custom model-training component. The remainder of this report reflects this adjusted approach: reviewing relevant prior work, detailing the methods for obtaining and analyzing AI-generated music from established models, presenting findings on their musical structures, and discussing what this implies about AI’s understanding of music theory. These insights are also connected to broader class themes around creativity and machine intelligence, and the quality of the resulting musical artifact is evaluated.

Related Work

AI-based music generation has evolved through several paradigms, ranging from rule-based systems to deep learning models [2]. Early algorithmic composition often relied on explicit music theory rules or probabilistic grammars to ensure valid output (for example, rule-based harmonizers for Bach chorales). As machine learning advanced, models began training directly on musical data, allowing systems to learn musical patterns implicitly. Below is a summary of key developments and pertinent literature:

Recurrent Neural Networks (RNNs)

These were among the first deep learning models for music. Performance RNN [4] is a notable example, using an LSTM to generate solo piano performances with expressive timing. It was trained on real piano competition data to capture nuances like velocity and tempo variations. While the output can sound quite realistic in short spans, RNNs often compress context into a fixed memory, leading to a loss of thematic continuity over longer durations. Other RNN-based systems, such as DeepBach [5], specifically targeted stylistic tasks (four-part chorale harmonization) by incorporating some music theory constraints, achieving high-quality results in that niche.

Transformers and Self-Attention

The introduction of self-attention mechanisms transformed sequence modeling, including music generation. Huang et al. (2018) applied Transformers to music in the Music Transformer, showing that relative attention enabled the model to capture periodicity and repeat motifs over long sequences [6]. Trained on the MAESTRO piano dataset, Music Transformer achieved state-of-the-art continuity and generated multi-minute pieces without losing thematic consistency. Around the same time, OpenAI's MuseNet extended the Transformer approach to a massive corpus of MIDI files from various genres, generating multi-instrument compositions of up to four minutes by predicting one token at a time in an autoregressive manner [7]. Notably, MuseNet was not explicitly programmed with music theory; it learned harmony, rhythm, and style implicitly from data [7]. These Transformer models produce impressively coherent music compared to earlier RNNs, though they are not flawless. For instance, Payne (MuseNet's creator) observed that occasional mistakes occur and that careful prompt design is needed to guide the style [2].

Specialized Models and Constraints

Some systems focus on narrower tasks or incorporate knowledge to ensure musical validity. Google's Bach Doodle (2019) built on a model called Coconet, trained on 306 Bach chorales, to harmonize short melodies in Bach's style [8]. Coconet learns to fill in missing notes in a musical score, making it adept at maintaining counterpoint rules and generating four-part harmonies from partial inputs [8]. The Bach Doodle interface allowed users to interact with the model, illustrating a user-friendly application of AI in music composition. Such style-specific approaches, including evolutionary algorithms for Bach chorales, demonstrate that AI can achieve near-flawless music-theoretical correctness when operating within a constrained style [2]. However, these systems may be less flexible in other genres or more experimental forms of music.

Audio-Based Generators

Another branch of research focuses on generating music as audio rather than symbolic notes. OpenAI's Jukebox exemplifies this approach by modeling raw audio, including vocals, at the waveform level. Jukebox can mimic certain artists or genres, although audio models often grapple with coherence and fidelity challenges, such as audible artifacts [2]. More recently, Google introduced MusicLM (2023), a model capable of generating high-fidelity music based on textual descriptions [6]. MusicLM follows a hierarchical strategy, first creating a high-level sequence (rough melody/chords) and then refining it into audio, which supports consistency over longer pieces of music. However, since it operates in the audio domain, analyzing its "understanding" of music theory requires transcription or careful listening.

Evaluation of AI Music

Evaluation remains a prominent theme in the literature. Civit et al. (2022) emphasize the complexity of assessing musical outcomes, including considerations of musicality (how pleasing it sounds), structure (how coherent it is), and genre-specific style or rule adherence. Many studies use listening tests or feedback from musicians; others rely on objective metrics such as counting rule violations in Bach chorales, examining tonal tension curves, or using music21 to detect chords and keys. Although Transformer-based models have significantly improved the perceived structure of generated music, they are often judged by listener preferences or basic theory correctness. A comprehensive quantitative measure of "musical quality" remains elusive.

In summary, previous research has established a range of models capable of generating music with varying degrees of musical understanding. Transformers like MuseNet and Music Transformer dominate the state of the art in symbolic music generation, enabling longer and more complex compositions. Meanwhile, style-specific systems such as Coconet leverage targeted data to excel in narrow tasks (e.g., Bach chorales). Building on these developments, the following study compares outputs from multiple systems—general-purpose vs. style-specific and symbolic vs. audio—to understand how design choices affect musical structure in the generated pieces.

Methods

Original Plan: Magenta Model Training and Theoretical Metrics

The original methodology involved training a deep learning model on a large piano performance dataset and then evaluating the generated music against music-theoretical criteria. To achieve this, the MAESTRO dataset (Hawthorne et al. 2018) was selected as the training corpus, given its size and quality. MAESTRO consists of approximately 200 hours of virtuosic piano performances captured as MIDI with aligned audio. The repertoire ranges from Baroque to early

20th-century works and provides fine temporal alignment, making it ideal for training an expressive performance model.

Data preparation began with downloading and preprocessing the MAESTRO dataset (v3.0.0). This included extracting MIDI files and potentially augmenting or filtering them. Although MAESTRO's rich performance data (such as pedal information and expressive timing) is valuable, for theory analysis it seemed practical to simplify to pitch and rhythm. This might involve quantizing the MIDI or focusing on note on/off and velocity data only.

The plan next called for model training using Google Magenta's implementations of Performance RNN and Music Transformer. Performance RNN (an LSTM-based model) is known for capturing expressive timing, whereas the attention-based Music Transformer excels at learning long-range musical structure. The idea was to establish a baseline with Performance RNN, then move on to Music Transformer for improved coherence. Training would involve configuring model hyperparameters, setting up a TensorFlow environment, and running multiple epochs on cloud-based hardware to accommodate the data size and model complexity.

Upon completing model training, the goal was to generate a set of new piano compositions—specifically, 5–10 pieces of a few minutes each. To maintain consistency, the plan included seeding the generation with a fixed primer melody or chord (for instance, a C major arpeggio), allowing a comparison of how different models would expand from the same starting point.

For chord progression analysis, the music21 library would be used to "chordify" the generated pieces, reducing each measure or beat to a chord symbol. This approach helps identify common harmonic patterns such as ii–V–I progressions or circle-of-fifths sequences. It also facilitates detecting dissonant or non-functional chords, as well as classical cadences (V–I, V–I with a leading tone, etc.) at the ends of phrases or entire pieces.

Key consistency was another important factor. Using music21's key analysis, the plan was to investigate whether each composition remained in a single key or modulated to others. Particular interest lay in whether any modulations followed logical pathways (like moving to the dominant or relative minor) or if the model modulated erratically, indicating a lack of tonal cohesion.

Melodic structure analysis aimed to identify repeated motifs or recognizable phrase structures. One question was whether the model would return to a theme (e.g., an ABA form) or develop motifs over time. Simple statistical measures—such as the distribution of interval sizes—could reveal whether the generated melodies favored stepwise motion, which human-composed melodies often do. There was also a plan to look for exact or transposed repeats of note sequences.

Finally, rhythmic and dynamic patterns were set for evaluation, especially regarding expressive performance. Although more aligned with performance than composition, it would be informative to see whether the model (particularly Performance RNN) inserted expressive timing and dynamics in ways consistent with musical phrasing, for example slowing down at cadences or emphasizing downbeats. This involves examining MIDI velocity patterns and note lengths.

Midway through the project, persistent technical roadblocks appeared while attempting to run Magenta's environment under Visual Studio Code. Build issues and dependency conflicts (for example, TensorFlow and CUDA version mismatches, plus older library requirements for Magenta) significantly impeded progress. Given the project timeline and the April 3 draft due date, a decision was made to change course. Instead of training a model from scratch, the plan shifted to analyzing music outputs from existing AI music models that are recognized as state-of-the-art. This adjustment allowed for continued exploration of AI-generated music without the overhead of training, though at the cost of having less control over the model architecture and training data. The expectation, however, was that these high-quality models would produce outputs suitable for meaningful analysis.

Three AI music generation systems were chosen to represent a range of capabilities and approaches:

OpenAI MuseNet (2019): MuseNet is a 72-layer transformer model trained on an extensive MIDI dataset encompassing classical to pop music (MuseNet | OpenAI). It can generate multi-instrument compositions and blend styles (e.g., Mozart in a jazz band style). For this analysis, the focus was on MuseNet's ability to generate solo piano pieces in a classical style, which facilitates direct assessment of tonal harmony and form. Access to MuseNet was obtained through an interactive tool provided by OpenAI, allowing a choice of style and optional composer or primer melody. Multiple samples were generated in "Classical Piano" mode, one prompted with a few notes in C major for anchoring and another started from scratch. The resulting MIDI files were saved for analysis.

Google Bach Doodle (Coconet, 2019): This model is a specialist system designed specifically to emulate Bach's four-part chorales. Via the archived Bach Doodle interface, a custom melody was entered in the soprano voice (eight notes over two measures in a simple C major theme), and the AI harmonized it. The Doodle interface allows downloading the resulting 4-voice MIDI output—Soprano (possibly slightly adjusted from the original input), Alto, Tenor, and Bass. This short extract is rich in vertical harmony, making it ideal for examining voice leading and chordal correctness. Since Coconet was trained extensively on actual Bach chorales, its harmonizations should adhere closely to rules of tonal harmony (avoiding parallel fifths, resolving leading tones properly, etc.).

Google MusicLM (2023): As a cutting-edge model that creates music from text descriptions, MusicLM makes it possible to test AI music generation in the audio domain. One of the example text prompts from the MusicLM demo page was used: "A calming piano solo, with a gentle melody that gradually evolves and creates a sense of resolution by the end." MusicLM then generated about 30 seconds of audio matching that description. An offline tool (AnthemScore) was used to transcribe the audio to MIDI, providing a rough approximation for symbolic analysis. While transcription is imperfect (particularly for polyphonic passages), it enabled a basic examination of notes and chords. Additionally, the audio was evaluated by ear to capture details that a MIDI transcription might miss, such as timbre or subtle performance nuances. Because MusicLM works in the audio domain, it offers insights into both structural elements (like melody/harmony) and expressive qualities (such as rubato), though the focus remained on theoretical characteristics.

With these models and outputs selected, the analysis pipeline largely mirrored the original plan but was tailored to each case. The music21 library in a Python environment parsed the MuseNet and Bach Doodle MIDI files, enabling access to note, chord, and key information. For each piece, the likely key (and any modulations) was computed, followed by an extraction of the chord sequence. In the Bach chorale, music21's roman numeral analysis was applied within the identified key to verify whether the model adhered to standard Bach progression patterns. For the lengthier MuseNet piece, each section (usually every 8 or 16 measures) was checked for key stability or modulations.

A custom script detected cadence points in the MuseNet and Bach outputs, using the heuristic that a cadence often involves a V (dominant) chord resolving to I (tonic), especially at a phrase boundary (marked by a rest or the end of the piece). In the Bach harmonization, a final authentic cadence was expected at the piece's conclusion. In MuseNet's piece, attention was paid to the ending, and any mid-piece pauses to locate cadential structures.

Melodic structure was also examined, particularly in MuseNet's freeform piano composition. The top note of the texture was extracted as the melody, and a search for repeated motifs (5–8 notes) was conducted. The melodic contour such as an arch shape and intervallic leaps were noted. In the Bach piece, the given melody formed the soprano voice, so the focus turned to the motion in the other voices—mostly stepwise or small intervals, aligning with chorale style guidelines.

Each model's output required slightly different handling—Bach Doodle provided a strict SATB format, MuseNet offered a more loosely structured piano texture, and MusicLM demanded transcription before symbolic analysis. Even so, using music21 as a core tool allowed a relatively consistent framework for investigating harmony, key, melody, and cadences. By comparing the findings from these diverse outputs, it became possible to conclude how each AI model addresses music theory concepts and structural coherence. All analysis code and relevant data

(MIDI files, scripts) have been documented, while this report focuses on presenting the methods and summarizing key observations.

Results

After collecting musical outputs from three AI systems and analyzing them, the findings are presented in terms of harmonic progressions, cadential structures, key/mode consistency, and melodic organization. The results are organized by model, highlighting the characteristics of each generated piece and noting the extent to which they align with or deviate from traditional music theory expectations:

MuseNet (Transformer, multi-style):

The MuseNet-generated piano piece (approximately three minutes, prompted in C major) stayed largely within a tonal center. Music21 analysis identified the key as C major for roughly the first 16 measures, followed by a brief modulation to G major (the dominant) for about 8 measures, before returning to C major. This mirrors common classical practice of modulating to the dominant in middle sections, suggesting the model learned a realistic pattern. The chord progressions were generally logical; for instance, one excerpt showed a sequence (Am – D7 – G – C), corresponding to a ii–V–I in G leading back to C (a secondary dominant approach). These progressions indicate the model absorbed some classical harmonic moves.

At least two authentic cadences were observed: one halfway through (ending on a G major chord, functioning as a half-cadence in the home key) and a final cadence ending on C major. The final cadence was a V–I (G to C) with the leading tone resolving correctly to the tonic in the melody. However, MuseNet was not perfect. A few places featured murkier harmony, such as a sudden chromatic chord (e.g., an unexpected E♭ major) that briefly created a modal mixture effect. These moments were rare and could be interpreted as creative leaps or minor lapses in tonal consistency.

In terms of melodic structure, the piece had a singing right-hand melody that repeated a turn-like figure motif, especially in the first section. The same motif reappeared near the end, providing a sense of closure. That said, the piece lacked a strong overarching form (such as a defined ABA structure), feeling more like a continuous rhapsody. MuseNet’s strength lay in local coherence, moment-to-moment; it sounded convincing with fluent voice leading and no glaring parallel or spacing errors. Yet on a global scale, there was some meandering. A listener might enjoy the pleasant harmonies and stylistic authenticity but find the piece somewhat aimless if a clear thematic development is desired.

Bach Doodle (Coconet, Bach Chorale style):

The Bach harmonization task output was a 4-part chorale in F major (the input melody was also in F major). As anticipated, the harmonization was highly tonal and rule-conforming. Each chord could be explained by common-practice harmony: the progression started on I (F major), moved through a ii–V–I cadence in the middle (G minor → C7 → F), and eventually ended on a perfect authentic cadence (C7 resolving to F major, with the soprano ending on A, the chord's third, for a final Picardy-third-like effect as the melody closed on the third).

Voice leading was very smooth. Music21 analysis did not flag direct parallel fifths or octaves between the voices. Alto, Tenor, and Bass parts generally moved by step or small leaps, in keeping with Bach's style; when larger leaps appeared, the lines changed direction to maintain balance. The chordal analysis showed use of secondary dominants: for example, an E7 chord (V/vi) led to a brief tonicization of D minor (vi). This indicates a strong grasp of functional harmony, likely internalized from the training chorales. The final cadence featured the bass moving from C (dominant) to F (tonic), and the leading tone E in the soprano resolving upward to F, forming a textbook authentic cadence.

Given the piece's short length (two phrases), long-term structure was not especially relevant, though each phrase had a clear beginning and ending marked by cadences. Qualitatively, the output could be mistaken for a snippet of an actual Bach chorale, except for a certain "safe" quality. It adhered closely to the rules and lacked any of Bach's more inventive melodic flourishes. There were no glaring errors, which underscores the model's strong grasp of Bach-style harmony. Overall, the Bach Doodle output demonstrated that a constrained objective, combined with robust training on a specific style, can yield near-perfect music-theoretical correctness. The limitation is domain specificity: it excels at Bach chorales but would not be suitable for freer forms or more modern harmonies.

MusicLM (Text-conditioned, audio model):

The piece generated by MusicLM for the prompt lasted about 30 seconds and featured a solo piano performance. The style matched the prompt's descriptors of "calming" and "evolving melody." From a music theory perspective, the transcription suggests the piece stayed in A minor (or An Aeolian). A repeating left-hand arpeggio pattern outlined the chord tones of A minor and D minor, implying a i–iv oscillation. Over this, the right hand played a simple melody that gradually rose in pitch, imparting a sense of development.

Despite the prompt mentioning "resolution by the end," the piece did not conclude with a classical V–I cadence. Instead, it finished on an A minor chord following a brief pause—a resolution of sorts (tonic chord) but without a preceding dominant. This is not out of place in a gentle, ambient piano style, where plagal or modal cadences (iv to i) are common, rather than a strong dominant-tonic cadence. The music remained tonally consistent; no sudden modulations or out-of-key chords were introduced, which contributed to the calm ambience.

Because MusicLM is an audio-generating model, the timbre and expressive continuity were notably strong. The piano tone sounded rich, and a slight tempo rubato effect added expressiveness. The melodic content was somewhat repetitive, featuring a short motif transposed to different pitch levels. Since MusicLM does not directly output notation, theoretical analysis required transcription and listening. The result was pleasant and tonally sound, with no obvious dissonances or adventurous harmonic moments likely due to the prompt specifying a safe, caring style rather than complex chord progressions. The piece's 30-second duration limits long-term structural evaluation, though it did feature a gentle build and a sense of winding down at the conclusion. From a listener's perspective, the output was coherent and satisfying. Lacking a strong cadence or modulation is not necessarily a flaw; it might reflect the model's exposure to a wide range of piano styles, including those with looser cadential requirements.

Table 1: Qualitative Comparison of Generated Music Outputs

Model & Output	Key Consistency	Harmonies & Cadences	Melodic Structure	Notable Limitations
MuseNet (Classical piano, ~3 min)	Mostly diatonic in C major; one brief logical modulation to G major (dominant).	Functional chords (e.g. ii–V–I progressions) present; contains a mid-piece half-cadence and a final V–I cadence in C.	Recurring motif provides some unity; lacks a clear large-scale form (through-composed feel).	Occasional odd chord (one chromatic out-of-key chord); overall direction is wandering without a strong theme development.
Bach Doodle (Chorale, 8 bars)	Completely diatonic in F major; any tonicizations are brief and within Bach style.	Every chord follows tonal harmony; it ends in a perfect authentic cadence (V^7–I in F) with proper voice-leading.	Constrained by input melody, output adds harmonies that follow standard chorale phrase structure.	Very short length: creativity limited to harmonization (no original melody composition); strictly Baroque style.
MusicLM (Calming piano, 30s audio)	Consistent A minor/A Aeolian throughout; no modulations.	Simple, consonant harmony (i–iv–v patterns); ends on tonic chord (gives sense of closure, though not a V–I cadence).	Evolving melody with slight variations on a motif; has a gentle rise and fall dynamic shaping a mini arc.	Hard to verify theory due to audio nature; harmony is simplistic (few chord changes); repetitive motifs risk sounding looped.

Interpretation

All three models avoided jarring musical mistakes and produced internally coherent output within their respective styles. MuseNet and Bach Doodle, both operating in the symbolic domain, demonstrated clear evidence of having learned tonal theory—incorporating cadences and appropriate chord progressions—while Bach Doodle, in particular, achieved near human-level harmonic correctness in chorale writing. MuseNet showed strong competency with only occasional lapses, possibly because its scope is broader (longer pieces, multiple styles). MusicLM's output, although analyzed with less precision due to its audio-based nature, remained tonally consistent and pleasant, suggesting that it did not generate out-of-key or random notes. This is likely the result of extensive exposure to harmonious examples during training, leading to the internalization of those statistical patterns. However, the absence of a pronounced cadential formula in MusicLM's piece highlights that certain musical constructs (like classical cadences) may not appear unless they are prompted or contextually required; the model prioritized fulfilling the prompt ("calming, evolving") rather than adhering to strict theoretical norms.

Discussion

The comparative analysis of MuseNet, Bach Doodle, and MusicLM outputs provides insight into the capabilities and limitations of AI models in mastering music theory concepts. Several key themes emerge from these results:

Implicit Theory Learning

A central question is whether AI models exhibit an understanding of music theory. The findings suggest that implicit learning from data can indeed yield behavior that mirrors theory compliance. For instance, MuseNet was never given explicit harmony rules, yet produced valid chord progressions and inserted a classical cadence. This implies that statistical learning captured essential aspects of tonality—essentially discovering the same patterns that music theory codifies by being exposed to many examples. Bach Doodle's success partly reflects the abundance of training data in a narrow style, but it is notable that it can generalize to harmonizing new melodies (unseen during training) while following learned rules. This indicates some degree of generalization: instead of memorizing exact pieces, the model applies stylistic constraints to new inputs, which suggests a form of understanding.

Nonetheless, caution is warranted: does the AI "understand," or does it simply replicate statistical patterns? For example, if MuseNet occasionally produces a strange chord, it does not "know" it broke a rule; it may not have had enough training examples for that particular context. These models do not consciously grasp "V chord resolves to I," yet the behavior emerges from observed data regularities. In essence, the knowledge of theory is tacit. It functions like a savant who can play by ear and often follows the rules without formal study. This works well until unusual circumstances arise, where explicit knowledge of the underlying rules could help. Minor evidence of this limitation appears in MuseNet's odd chromatic chord, which a theory-aware composer might use more deliberately or resolve more smoothly.

Model Constraints and Musical Form

The design and training of each model imposes certain constraints on musical output. Bach Doodle's limited output length and specific style guarantee clear cadences and phrases, mirroring its training examples. MuseNet, having absorbed data from an array of forms (symphonies, pop songs, etc.), generates a more "average" sense of structure, offering no guarantee of prominent section breaks. Meanwhile, MusicLM, guided by a text prompt, is expected to adhere to the specified structure or mood. When a prompt implies a single mood, the model sustains it throughout.

This suggests that external constraints or guide rails sometimes help produce more structured output. For example, prompt engineering for MuseNet can specify a particular progression or chord sequence, or a rule-based post-processing step can ensure the piece ends on a tonic chord. Some research merges AI generation with rule-checkers to correct mistakes, but this can reduce creativity. The analysis indicates that these models, left on their own, often default to "safe" patterns learned from data. While these patterns are mostly correct, they do not necessarily exhibit innovative forms. Rarely do they surprise with a clever modulation or extensive thematic recall, since such events are statistically less common or demand an intentionality that the models do not possess.

Musicianship and Expressiveness

Although the focus here is on theory, musical elements beyond raw notes—dynamics, timbre, and expression—are also relevant. Performance RNN's emphasis on expressive timing [1] and MusicLM's high-quality audio reveal that performance aspects are increasingly addressed by AI. In the MusicLM example, rubato contributed to the musicality. A composition may be theoretically correct but still feel mechanical if lacking human-like expressiveness. AI that generates only MIDI data may not capture nuances of performance as readily, whereas audio-focused AI often embeds some expression.

Consequently, "understanding music" involves more than just placing the correct notes in sequence. Phrasing, dynamics, and interpretive choices are integral. The models discussed generally do not explicitly dictate phrasing or dynamics (Performance RNN does, but it was not run in this particular analysis; MuseNet and Bach Doodle outputs mostly feature flat dynamics unless present in the training data). For a fully realized performance, it might be necessary to rely on either post-processing or a human performer. Theory and expression are intertwined in music; while AI's adherence to theoretical rules is part of the picture, emotional expressiveness remains an area where human musicians often retain an edge, even though AI is improving in that domain.

Ensuring Fairness and Values in Creative AI Systems

Switching to an evaluative study also opened up space to consider fairness, justice, and values in AI music – themes that might be sidelined in a pure model-building exercise. A key lesson from the course is that AI systems are never neutral; they carry the values (or biases) of their training data and design. Data feminism principles remind us that *"data are not neutral or objective. They are the products of unequal social relations,"* and that context is essential for ethical analysis [4]. Applying this lens, we asked: Whose music and theory does MuseNet or MusicLM implicitly prioritize? Are certain genres or cultural expressions underrepresented? Our evaluation criteria were implicitly value-driven to catch such issues. If a model persistently violates basic music theory like producing dissonant intervals or nonsensical chord progressions, it could indicate either a data gap, perhaps training data lacking those theory-rich examples or a design that doesn't honor known musical *values* like harmonic consonance. By critiquing these outputs, we were engaging in a form of value-sensitive evaluation, examining whether the AI's behavior aligns with human musical.

Fairness and justice in AI were broader class themes that we also connected to our project. Fairness in a creative AI context can be interpreted in multiple ways. One aspect is representation fairness: does the AI treat different musical traditions equitably? We reflected on whether systems like MusicLM, trained on 280,000 hours of audio, might over-represent Western popular and classical music while under-representing indigenous, folk, or experimental music. This touches on *data justice*: if an AI music model largely generates in the style of Western composers, it risks reinforcing a single cultural aesthetic as "default," marginalizing other voices. Our awareness of this issue was heightened by readings on fairness in ML that discuss how unequal training data leads to unequal performance. For example, the *Fairness and ML* textbook notes that models can have "terribly [high error] for a minority group" if that group's data is scarce [7]. In music generation, "error" might translate to stylistic blindness – an inability to authentically produce a less-seen style. Although we did not quantitatively measure genre coverage, our critical listening kept this concern in mind.

Another aspect of fairness is who benefits or is harmed by AI music systems – a question of justice and power. Value-sensitive and feminist frameworks push us to ask: are these tools empowering musicians and audiences, or mainly benefiting tech companies? We considered, for instance, that an AI able to imitate living composers could raise fairness issues if those composers neither consented nor were credited. This is analogous to image-generation AI controversies, but in music, the debate is emerging: is it just to use an artist's catalog to train a model that might then undercut the need for that artist's work? While our project focus was on output quality and characteristics, these ethical questions hovered in the background as part of the context we discussed. Ultimately, integrating fairness and value frameworks meant that our evaluation was not just about which model performs best, but also about which model's outputs align with human values of inclusivity, respect, and musical authenticity. By explicitly referencing these concerns in our report, we demonstrate a critical perspective: the quality of AI music is not only a technical matter, but also a question of how well the system's behavior aligns with diverse human values and equitable treatment of musical cultures.

Ethical Data Practices and Governance in Music AI

A significant theme that emerged during this project was data governance, encompassing data sharing, ownership, and the consideration of indigenous data principles. Music data introduces a distinct set of ethical and legal challenges. Initially, the plan involved collecting or using a dataset for training. Perhaps MIDI files of classical music or folk songs. Early on, however, it became evident that issues of ownership and permission could be complex. Who owns these compositions, and is it permissible to use them for AI modeling? What if traditional or Indigenous music is included? Would that amount to cultural appropriation for a tech endeavor? Such questions parallel those raised by Carroll et al. (2020) on Indigenous data sovereignty. The CARE Principles (Collective Benefit, Authority to Control, Responsibility, Ethics) emphasize that data involving Indigenous peoples should be used in ways that honor their rights and interests, rather than being freely open-sourced. Although this project did not rely on Indigenous data specifically, the principle remained relevant: data is not merely a public resource to extract; ethical usage requires respecting the communities and creators behind it. This perspective also influenced the decision to pivot from direct model training. One reason for choosing to evaluate Google's and OpenAI's systems instead was to avoid engaging in potentially problematic data collection without proper governance. At the same time, this shift did not resolve data issues entirely; it merely changed the focus to critiquing the data practices of these companies. When examining how MuseNet and MusicLM were trained, mostly on publicly available music (e.g., MIDI archives or possibly YouTube audio), questions arose about transparency and copyright. Google's MusicLM paper mentions an extensive dataset but lacks detail about precisely which music was used or how it was obtained. This opacity is itself a governance concern, highlighting the tension between open data for scrutiny and potential violations of intellectual property rights.

The concept of data context and power from Data Feminism also played a role. A key assertion is that data analysis must account for context and the unequal power relations inherent in data collection. Applied here, it became clear that the training data for AI music systems does not exist in a vacuum. It reflects power imbalances in how music is produced, distributed, and documented. Western music is abundantly notated in MIDI or made digitally accessible, whereas many non-Western musical traditions remain underrepresented online. As a result, relying on "available" data can entrench a cultural bias favoring Western music. This extends beyond legal permissions to broader questions of whose data is considered valid or important. If MusicLM relies on a massive but skewed dataset, its impressive output may inadvertently erase minority traditions by failing to generate music that it was never exposed to. This observation stems from the idea that data can be a form of global power, allowing those with large music datasets often tech giants to shape what AI deems "standard" knowledge. The CARE principles recognize the tension between open data and community rights, prompting reflection on whether training with public domain folk recordings genuinely benefits source communities or primarily serves corporate users. The CARE ideals (Collective Benefit, Authority, Responsibility, Ethics) often clash with the drive to employ FAIR data (Findable, Accessible, Interoperable, Reusable) in AI systems that focus on whatever is readily available. In reality, many contemporary AI music systems prioritize FAIR and pay less attention to CARE, and part of this project's goal was to highlight that imbalance and suggest that more inclusive, value-sensitive data practices could be applied in future work.

Lastly, intellectual property and authorship concerns also fell under the purview of data governance. MusicLM reportedly remained unreleased, in part, due to "training data memorization," where it would occasionally reproduce exact fragments of existing songs raising both ethical and legal red flags. This real-world example underscores the downstream importance of governance. Without safeguards, an AI might inadvertently produce someone else's creative work without attribution, violating norms of ethical data usage. This aligns with the notion of Responsibility and Ethics in the CARE framework. Consequently, a responsible evaluation of AI music systems should include checking for potential plagiarism or overfitting. Had the generated outputs contained recognizable sections from existing compositions, it would have signaled a breakdown in governance during the training process. In these ways, data governance considerations permeated every stage of this project: from deciding not to gather potentially copyrighted music ourselves, to examining whether large-scale models appropriately respected the boundaries of their training data. By weaving these concerns into the project's analysis, it was possible to connect day-to-day decisions with broader principles of data governance and ethics covered in class.

Creativity, Authorship, and Authenticity: A Reflective Synthesis

Connecting these class themes to the project added depth to the discussion of creativity, authorship, and musical authenticity in the age of AI. By examining human-centered design, fairness, data governance, and data infrastructure, the investigation moved beyond a purely technical lens to ask, "What does it mean for an AI to create music?" The pivot to evaluating existing models, informed by these frameworks, enabled a view of AI outputs not merely as finished compositions, but as artifacts of a socio-technical system that warrants critical scrutiny.

A central debate involved whether these systems displayed genuine creativity or merely reassembled patterns from vast training data. The conclusion leaned toward a hybrid understanding. AI models do leverage existing patterns (data-driven creativity), but sometimes their recombination or emergent musical structures surprise even experienced musicians, hinting at forms of computationally augmented creativity. This view aligns with human-centered AI ideals: the notion that AI should "augment and enhance humans' abilities" rather than displace them [5], [11]. For example, one could imagine a composer collaborating with MusicLM for initial melodic ideas, then weaving in personal expression and narrative to shape a final piece. Such co-creative scenarios frame the AI as a tool for extending, rather than replacing, the creative process.

Musical authenticity was another recurring theme, investigated through a human-centered design lens. Authentic art typically emerges from human intention and context. While AI can mimic stylistic elements and evoke emotional responses, there is often a subtle sense—particularly among trained listeners—that something is missing. It may be the risk-taking or personal narrative that derives from lived human experience, an element not captured by raw data. Data Feminism's call to "consider context" [4] is especially relevant here: AI lacks the lived experiences that fuel human artistry. On one level, an AI system might achieve surface-level authenticity (correct notes and stylistic fidelity), but it can struggle to convey deeper cultural meaning, particularly if it does not understand the human motivations or cultural contexts behind its training material. At the same time, humans often operate implicitly in their own musical traditions: a blues guitarist might not articulate theory formally but still create an authentic sound through exposure and practice. In this sense, an AI trained on large amounts of blues can similarly learn to replicate patterns without formal theoretical knowledge, blurring the line between "authentic" and "derivative."

Overall Artifact Quality

The "artifact" produced and analyzed in this project is two-fold: the collection of AI-generated musical pieces (from MuseNet, Bach Doodle, and MusicLM), and this analytical report itself. We assess the quality of each, acknowledging that this is a draft stage.

The AI-generated music we obtained can be considered the core creative artifact. Overall, the quality of these pieces is surprisingly high given no human composed them outright. The Bach Doodle harmonization stands out for quality – if one judges it by the standards of four-part harmony, it scores near top marks (correct voice-leading, clear cadences, rich inner parts). As a chorale, it is short and perhaps unremarkable (one could say it's a perfectly "competent" piece in Bach style, though maybe lacking Bach's personal flair). Still, as an artifact, it serves its purpose excellently and could even be used in an educational context (e.g., a music teacher could show it as a correct harmonization for a given melody). The MuseNet composition is a bit harder to rate, as it is longer and more free form. In terms of aesthetics, it is pleasant to listen to – it arguably could pass as a relaxing improvisation by a skilled pianist who's riffing in classical style. It has no glaring mistakes that would make a listener cringe. However, from an artistic standpoint, it might be judged as mediocre music: it doesn’t have a memorable theme or a strong emotional arc; it’s the kind of piece that might serve as background music rather than command attention on a concert stage. That said, considering it was generated by AI, the coherence and style match are impressive. Many listeners might not realize it was machine-made unless told. The MusicLM audio piece was also of good quality: the piano sound was realistic, and the composition was soothing and on-point with the prompt. Its quality as an artifact is slightly marred by its brevity and slight repetitiveness – it felt more like a demo or sketch than a fully fleshed-out piece. If we compare it to production music (like stock music one might license for a video background), it meets that bar: it’s tonally pleasant, non-intrusive, and sets a mood. In terms of musicality, one could rate it as decent but not exceptional; as a proof of concept of AI abilities, it is excellent.

Reflecting on the combined artifact (the music + the analysis), a unique aspect is that the music itself can be considered a demonstration of the analysis points. In a final submission, we might include musical notation examples or audio clips as figures to illustrate specific findings (due to the nature of this text report, we described them in words and a table). The quality of the artifact in an educational sense is good: anyone reading this report can learn about AI music capabilities and even get a sense of how to analyze music with music21. If the measure of quality is also how much *we*, the project team, learned or created knowledge, then this artifact has value: we documented a real exploration that yielded insights.

Finally, it’s worth noting that artifact quality in creative projects can also refer to originality and contribution. While our pivot meant we weren’t creating a novel AI model, we did create a comparative analysis that, in a small way, adds to understanding how different models stack up. There is originality in the approach of evaluating multiple systems with a theory lens in one report. The transparency about the process (including challenges) also adds to the artifact’s authenticity and educational value.

References

[1] "Performance RNN: Generating Music with Expressive Timing and Dynamics," Magenta. Accessed: Apr. 03, 2025. [Online]. Available: https://magenta.tensorflow.org/performance-rnn

[2] M. Civit, J. Civit-Masot, F. Cuadrado, and M. J. Escalona, "A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends," *Expert Syst. Appl.*, vol. 209, p. 118190, Dec. 2022, doi: 10.1016/j.eswa.2022.118190.

[3] "MusicLM - AI Model for Music Generation." Accessed: Apr. 03, 2025. [Online]. Available: https://musiclm.com/

[4] K. Navaroli, "Week 11 Notes and Reflection on Data Feminism," My Site. Accessed: Apr. 03, 2025. [Online]. Available: https://www.kdnavaroli.com/post/week-10-notes-and-reflection-on-data-feminism-1

[5] H. Zhu, B. Yu, A. Halfaker, and L. Terveen, "Value-Sensitive Algorithm Design: Method, Case Study, and Lessons," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, pp. 1–23, Nov. 2018, doi: 10.1145/3274463.

[6] "Music Transformer: Generating Music with Long-Term Structure," Magenta. Accessed: Apr. 03, 2025. [Online]. Available: https://magenta.tensorflow.org/music-transformer

[7] S. Barocas, M. Hardt, and A. Narayanan, "Fairness and Machine Learning".

[8] S. R. Carroll *et al.*, "The CARE Principles for Indigenous Data Governance," *Data Sci. J.*, vol. 19, no. 1, Nov. 2020, doi: 10.5334/dsj-2020-043.

[9] E. Bietti, "Data is Infrastructure," *SSRN Electron. J.*, 2025, doi: 10.2139/ssrn.5041965.

[10] Z. Hammoudeh and D. Lowd, "Training Data Influence Analysis and Estimation: A Survey," *Mach. Learn.*, vol. 113, no. 5, pp. 2351–2403, May 2024, doi: 10.1007/s10994-023-06495-7.

[11] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, "'Everyone wants to do the model work, not the data work': Data Cascades in High-Stakes AI," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, May 2021, pp. 1–15. doi: 10.1145/3411764.3445518.