Modeling 53-TET Chord Progressions with a Microtonal GPT-2

David Dalmazzo
Music Technology Group
Universitat Pompeu Fabra, Barcelona, Spain
david.cabrera@upf.edu

Ken Déguernel
Univ. Lille, CNRS, Centrale Lille
UMR 9189 CRIStAL, F-59000, Lille, France
ken.deguernel@cnrs.fr

Abstract

This work explores the generation of harmonic chord progressions in 53-tone equal temperament (53-TET) using a GPT architecture, conditioned on microtonal transformation type, musical style, formal structure, and tonality. The training corpus originates from *iRealPro* lead sheets parsed into symbolic chord sequences and rendered to MIDI-MPE with voicings. Data augmentation transposes each song through all 53 pitch classes and applies 14 modal transformations, each defined by a holdrian-comma interval vector that redistributes diatonic steps across the 53-TET continuum, producing distinct microtonal modes from seven base scale types. The resulting corpus comprises 672,840 paired files combining symbolic metadata with MIDI-MPE. Each chord is represented as two consecutive blocks within a single flat autoregressive stream sharing one vocabulary and one prediction head: a symbolic block (L1) encoding the chord label, duration, and surrounding form, and a realization block (L2) encoding its voicing as pitch–velocity tokens. We train and compare two variants that differ only in how L2 pitches are encoded: Model A emits MIDI notes with per-note MPE pitch-bend deviations that realize the 53-TET target directly as playable MIDI, while Model B emits pure 53-TET step indices over a holdrian-comma vocabulary and defers the translation to MIDI with pitch bend to a deterministic post-processing step at generation time. A four-dimensional psychoacoustic dissonance vector, characterizing each tetrad by the intervallic qualities of its third, fifth, and seventh together with an aggregate sensory dissonance value, is injected as a continuous positional embedding shared across all tokens of the chord.

1 INTRODUCTION

This paper investigates the use of a Transformer architecture to generate chord progressions in 53-tone equal temperament (53-TET), a tuning system that offers a finer harmonic vocabulary, including chord qualities, interval shadings, and voice-leading possibilities not found in 12-TET, as well as supporting creative agency in music creation (Wiggins, 2016; Nielsen, 2026). Three questions structure the work: How to construct a paired symbolic and audio dataset large enough to train a generative model in 53-TET? How to encode both the symbolic identity and the sounding realization of each chord within a single learned vocabulary? And do the resulting progressions sustain harmonic coherence when they include chord qualities unfamiliar to listeners trained on 12-TET music? The system is designed as a compositional tool that preserves the musician's creative agency (Déguernel et al., 2022), suggesting harmonic paths rather than producing finished artifacts.

The contributions of this paper are:

1. A synthetic microtonal dataset of 672,840 paired MIDI-MPE and symbolic chord-progression files in 53-TET, generated by applying 14 modal scale types to a corpus of popular-music lead sheets that were previously transposed across all chromatic keys.[1]
2. A four-dimensional psychoacoustic embedding derived from the Sethares (2005) dissonance model, that we call *Harmonic Eigenspace*, computed per chord and injected into the positional encoding of a GPT-2 architecture, providing the network with a continuous, transposition-invariant signal about the harmonic quality of each chord.[2]
3. A two-level tokenization in which the symbolic label of each chord (root, quality, extensions, form context) and its sound realization (53-TET pitches and velocities) occupy consecutive positions in a single autoregressive stream, letting the model learn the mapping from symbol to voicing as part of the same prediction task.[3]
4. A controlled comparison through a listening test between two pitch-token semantics (MIDI-derived vs. pure holdrian-comma indices), isolating the effect of the pitch vocabulary on the model's ability to generate chord progressions that include both familiar harmonic motion and chord qualities that have no 12-TET equivalent.

2 RELATED WORK

The field of harmony modeling using neural networks has grown substantially in recent years; two recent surveys (Le et al., 2025; Ji et al., 2023) map the territory in detail, and we do not attempt to reproduce that coverage here. Instead, we focus on the three strands most relevant to the present contribution: Transformer-based modeling of chord progressions, harmony-aware and structurally conditioned symbolic music generation, and the computational treatment of microtonal harmony.

[1] The dataset is available at https://zenodo.org/records/19915101

[2] https://dazzid.github.io/ANIMA_Harmonic_Eigenspace/

[3] https://github.com/Dazzid/ANIMA_Microtonal_GPT

2.1 Transformer Models for Chord Progressions

The application of sequence modeling to symbolic harmony with neural networks predates the Transformer, with early LSTM approaches trained on text representations of chord symbols derived from Realbooks and Fakebooks (Choi et al., 2016). The introduction of the Transformer architecture (Vaswani et al., 2017) shifted the field towards attention-based models capable of capturing longer-range harmonic dependencies. Wu and Yang (2020) trained the Jazz Transformer on the Weimar Jazz Database, producing progressions that integrate chord, melody, and sectional information but also revealing persistent weaknesses in long-term structural coherence. Subsequent work targeted harmonic analysis directly: Chen and Su (2021) adapted Transformers for symbolic chord recognition in classical repertoire, and Zeng et al. (2021) released MusicBERT, a BERT-style encoder pretrained on the Meta MIDI Dataset (Ens and Pasquier, 2021) for downstream tasks including accompaniment suggestion and style classification. Li and Sung (2023) proposed MrBERT for melody and rhythm pretraining followed by sequence-to-sequence chord generation, though constrained to a triadic vocabulary.

Transformer models trained directly on chord-progression data (Dalmazzo et al., 2024a,b) have explored alternative tokenization strategies, including compact tuple encodings that separate root, quality, and extensions, and voicing-aware extensions that attach a rule-based MIDI voicing to the predicted chord symbol. These models operate entirely in 12-TET; voicing, when present, is derived from the chord symbol by a deterministic algorithm rather than learned jointly with the symbol. More recent work has continued pushing the chord-aware direction: Zhu et al. (2024) introduced MMT-BERT, a chord-aware multitrack Transformer that uses a fine-tuned MusicBERT as a discriminator within a GAN framework, and Luo (2024) proposed a D_{12}-equivariant Transformer that encodes the dihedral symmetries of the 12-tone equal-tempered system directly into the attention mechanism for chord-progression accompaniment. Both lines of work remain within 12-TET.

2.2 Harmony-Aware and Structurally Conditioned Generation

A parallel concern in the recent literature is how to inject high-level musical structure into otherwise autoregressive models. The Pop Music Transformer (Huang and Yang, 2020) introduced a beat-based token representation that made bar and position information explicit, while the Theme Transformer (Shih et al., 2022) conditioned generation on a user-supplied theme to improve long-range thematic coherence. Guo et al. (2022) proposed a domain-knowledge-inspired music embedding space and an associated attention mechanism for symbolic music modeling, explicitly exposing musical relationships that the raw token sequence does not carry. In the audio domain, Lan et al. (2024) introduced chord and rhythm conditioning for a Transformer-based text-to-music system. Chord-conditioned symbolic generation remains an active research direction (Le et al., 2025; Ji et al., 2023), with harmonic coherence and long-term structural integrity recurring as challenges in these surveys.

Most of these systems either treat harmony as a conditioning signal on top of a note-level sequence or operate purely at the chord-symbol level, without a principled account of voicing. Neither approach directly couples the symbolic label of a chord to a concrete, instrumentally plausible realization within the same learned vocabulary.

2.3 Microtonality in Computation

Microtonal composition has a sustained computational history. Anders and Miranda (2010) developed a rule-based constraint system supporting multiple temperaments, and Hirai (2023) released a microtonal MIDI and audio dataset to support machine-learning research in this area. The theoretical grounding for 53-TET reaches back to Holder (1731) and was formalized by Fokker (1955, 1963), with Muzzulini (2020) documenting Newton's early investigations of equal divisions of the octave. 53-TET tempers out both the schisma and the kleisma, aligning simultaneously with Pythagorean and five-limit just intonation (Tanaka, 1890), and it remains the theoretical basis of Turkish makam music (Yarman, 2007). Hart (2016) surveys microtonal practice in electronic dance music, and Chadwin (2019) examines microtonality in pop songwriting. Interactive approaches to microtonal harmony have emerged alongside this theoretical work, including constraint-based composition environments and recent applications for 53-TET modal interchange (Dalmazzo et al., 2025). Performance hardware such as the Lumatone isomorphic keyboard (Innovations, 2024) has also been incorporated by microtonal practitioners. Despite this activity, no Transformer-based harmony model has been trained natively on microtonal chord progressions, and the creative space of 53-TET harmony, including chords and progressions that fall outside the 12-TET grid, remains largely unexplored by data-driven models.

3 NOTATION FOR 53-TET MICROTONAL HARMONIES

In 12-TET, intervals such as the third or seventh of a chord are classified as either major or minor. In 53-TET, the smallest step is a holdrian-comma, and each of these intervals admits finer gradations forming a spectrum of qualities principally named subminor, minor, neutral, major, and supermajor. These conventions are borrowed from 31-TET, which is closely related to 53-TET; in 53-TET, each of the five qualities admits further variants (e.g., downminor, upmajor), producing ten gradations of the third and ten of the seventh. The naming convention used throughout this paper classifies each chord interval into a semantic category based on its holdrian-comma step count, then combines abbreviations for the third, fifth, and seventh to produce the chord symbol.

Each chord symbol combines a triad prefix (from the third), a fifth suffix, and a seventh suffix. The triad prefix encodes the third quality: major is `M`, minor is `m`, and microtonal gradations use abbreviations such as `sm` (subminor), `vm` (downminor), `N` (neutral), `vM` (downmajor), `^M` (upmajor), and `S^` (supermajor). The fifth is empty when perfect; non-standard fifths append `dim`, `+`. The seventh follows the same gradation.

Table 1 shows the full 10×10 grid of 100 seventh chords produced by combining ten third qualities (rows) with ten seventh qualities (columns), all with a perfect fifth at step 31. The diagonal entries (`SMSM7`, `^M^M7`, `maj7`, `vMvM7`, `NN7`, `nn7`, `^m^m7`, `m7`, `vmvm7`, `smsm7`) are the chords where third and seventh share the same gradation. Of these, only three (`maj7`, `m7`, and the dominant `7` at row

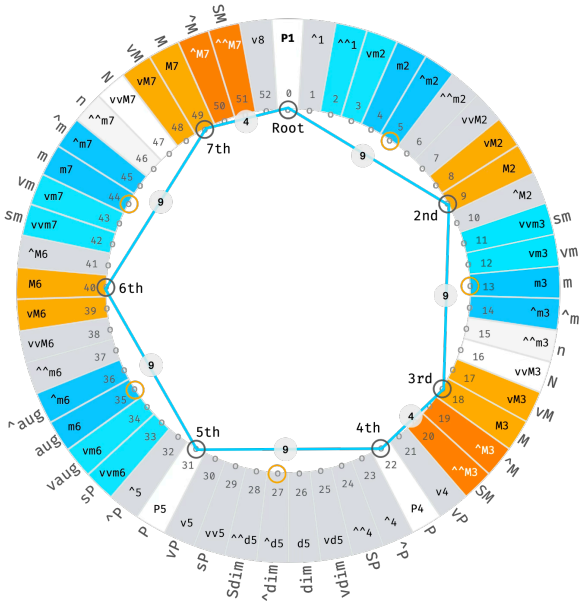

Figure 1: Circular representation of the 53-TET modal scale construction, with the scale formed out of seven notes and their related holdrian-comma distances.

M, column m) correspond to standard 12-TET qualities. The remaining 97 entries exist only in 53-TET.

Root names follow the same comma-level logic. Standard note letters with conventional accidentals (`#`, `b`) are extended with microtonal prefixes: `^` and `^^` raise the pitch respectively by one or two holdrian-commas, `v` and `vv` lower it by the same amounts. For instance, A chord symbol such as `^C vMvM7` denotes a downmajor seventh chord rooted one holdrian-comma above C.

4 DATASET

The training corpus originates from 4,005 popular-music chord progressions extracted from the *iReal Pro* application using the pipeline described in Dalmazzo et al. (2024a). Each piece is transposed to all twelve chromatic keys via a line-of-fifths algorithm that corrects enharmonic errors, producing 48,060 chord progressions in 12-TET (Dalmazzo et al., 2024b). For each song, two aligned files are generated: a MIDI file encoding the sounding pitches, durations, and velocities, and a text sidecar encoding the symbolic content that MIDI cannot carry (style label, tonality, form sections, repeat markers, and per-chord tuples of duration, root, quality, extensions, and slash chords).

To extend the corpus to 53-TET, we define 14 modal scale types, organized into seven base modes, each paired with its minor rotation. Every scale type is specified as a seven-element vector of holdrian-comma intervals that redistributes the 53 steps of the octave across the seven diatonic degrees (see Figure 1). Table 2 lists the seven base modes; the minor rotation of each is obtained by cycling the interval vector to start from the sixth degree.

Type 0 preserves the standard major scale intervals and acts as a baseline. The remaining six types redistribute the holdrian-commas, producing intervals that fall outside the 12-TET grid: type 1, for instance, narrows the major second and major third toward neutral intervals (neither major nor minor), while type 2 compresses the third to 3 holdrian-commas (a subminor third) and widens the fourth to 10. Each type, together with its minor rotation, reshapes the harmonic identity of every chord in the original progression. The familiar 12 pitch classes are repositioned within the 53-TET continuum: roots acquire comma-level shifts, thirds and sevenths move to new gradations, and the resulting chord qualities (vMvM7, NN7, smsm7, and others listed in Table 1) have no equivalent in the source 12-TET vocabulary. What was a standard chord progression in the original dataset becomes, under each transformation type, a progression through chord qualities native to 53-TET, preserving the melodic contour and formal structure of the original while inhabiting a different harmonic world.

For each of the 48,060 12-TET files, the conversion maps every MIDI pitch class to its 53-TET target under the selected scale type. The deviation between each target and its nearest 12-TET pitch is encoded as a per-note MPE pitch bend, so that every note in the output MIDI file carries its own pitch-bend channel. The text sidecar is simultaneously converted: chord roots and qualities are renamed to their 53-TET equivalents following a microtonal naming convention for 53-TET modal interchange (Dalmazzo et al., 2025). The convention classifies each chord interval (third, fifth, seventh) by its holdrian-comma distance into semantic categories (*e.g.*, subminor, neutral, upmajor for thirds), then combines abbreviations to produce the chord symbol. Root names extend standard note letters with microtonal prefixes (`^`, `^^`, `v`, `vv`) indicating comma-level pitch adjustments within 53-TET.

Applying the 14 scale types to the 48,060 base files yields 672,840 paired MIDI and text sidecar files. Of these, 322 files (23 base songs across all key transpositions) fail a chord-count alignment check between MIDI and text sidecar and are excluded, leaving 672,518 usable paired files.

5 HARMONIC EIGENSPACE AND POSITIONAL EMBEDDING

To encode harmonic quality beyond sequential position, we augment the positional embedding with a continuous four-dimensional vector $(\alpha, \beta, \gamma, D)$ that encodes the psychoacoustic profile of each chord, derived from what we call the *Harmonic Eigenspace*.

5.1 Harmonic Eigenspace

When two complex tones are sounded simultaneously, perceived roughness arises from interactions between their partials. Plomp and Levelt (1965) measured this effect for pure sinusoids and found that dissonance rises from zero at unison to a maximum near one quarter of the auditory critical bandwidth, then decays as the two tones become individually resolvable. Sethares (2005) fitted this shape with a difference of exponentials, scaled by the local critical bandwidth and weighted by the amplitude of the quieter partial (see Figure 2):

$$d(f_1, f_2, a_1, a_2) = a_{12}\left[e^{-b_1\, s\,(f_2-f_1)} - e^{-b_2\, s\,(f_2-f_1)}\right] \tag{1}$$

where $f_1 < f_2$, $a_{12} = \min(a_1, a_2)$, $b_1 = 3.5$ and $b_2 = 5.75$ are fitted constants, and s rescales the frequency axis to account for the critical bandwidth growing with frequency. For a complex tone with n harmonic partials, the

Table 1: 53-TET main grid: 100 seventh chords with perfect fifth (step 31). Rows are third qualities, columns are seventh qualities. The symbol is assembled as `{third prefix}{seventh suffix}`. Three cells correspond to standard 12-TET chords (maj7, 7, m7); the rest are microtonal.

	SM	^M	maj	vM	N	n	^m	m	vm	sm
SM	SMSM7	SM^M7	SMmaj7	SMvM7	SMN7	SMn7	SM^m7	SMm7	SMvm7	SMsm7
^M	^MSM7	^M^M7	^Mmaj7	^MvM7	^MN7	^Mn7	^M^m7	^Mm7	^Mvm7	^Msm7
M	MSM7	M^M7	maj7	MvM7	MN7	Mn7	M^m7	7	Mvm7	Msm7
vM	vMSM7	vM^M7	vMmaj7	vMvM7	vMN7	vMn7	vM^m7	vMm7	vMvm7	vMsm7
N	NSM7	N^M7	Nmaj7	NvM7	NN7	Nn7	N^m7	Nm7	Nvm7	Nsm7
n	nSM7	n^M7	nmaj7	nvM7	nN7	nn7	n^m7	nm7	nvm7	nsm7
^m	^mSM7	^m^M7	^mmaj7	^mvM7	^mN7	^mn7	^m^m7	^mm7	^mvm7	^msm7
m	mSM7	m^M7	mmaj7	mvM7	mN7	mn7	m^m7	m7	mvm7	msm7
vm	vmSM7	vm^M7	vmmaj7	vmvM7	vmN7	vmn7	vm^m7	vmm7	vmvm7	vmsm7
sm	smSM7	sm^M7	smmaj7	smvM7	smN7	smn7	sm^m7	smm7	smvm7	smsm7

Table 2: The seven base modal scale types. Each row shows holdrian-comma intervals between consecutive diatonic degrees (summing to 53). Each type has a corresponding minor rotation obtained by cycling the vector.

Type	Intervals distances	Description
0	9, 9, 4, 9, 9, 9, 4	Standard major
1	8, 7, 7, 9, 8, 7, 7	Neutral mode
2	9, 3, 10, 9, 3, 10, 9	Subminor mode
3	9, 8, 5, 9, 8, 5, 9	Harmonic 3rd & 7th
4	10, 9, 3, 9, 10, 9, 3	Up-major mode
5	9, 9, 5, 8, 8, 9, 5	downMajor 7th
6	8, 7, 7, 9, 5, 10, 7	Neutral N mode

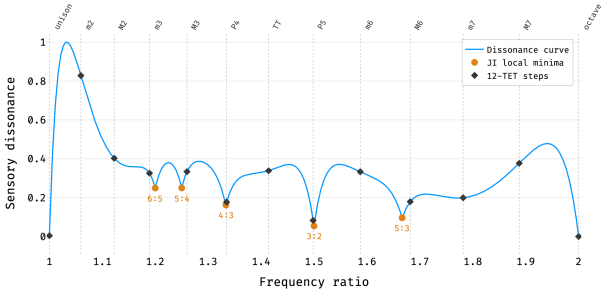

Figure 2: Spectral dissonance of the base frequency $D_F(\alpha)$ and the dissonance curves against the frequency ratio α over one octave. Local minima (orange markers) align with the just-intonation intervals. The 12-TET note references are aligned and shown with black diamonds.

total dissonance between two notes at a frequency ratio α is the sum of d over all pairs of interacting partials. Evaluated across one octave, $D_F(\alpha)$ produces a curve whose minima coincide with just-intonation intervals for harmonic timbres.

For a four-note chord, the root is fixed as the reference, and the three upper voices are parameterized by frequency ratios α, β, and γ relative to the root, all within one octave. The total dissonance is the sum over all six pairwise interactions, as shown in Figure 3:

$$D(\alpha,\beta,\gamma) = D_F(\alpha) + D_F(\beta) + D_F(\gamma) + D_{\alpha F}\left(\frac{\beta}{\alpha}\right) + D_{\alpha F}\left(\frac{\gamma}{\alpha}\right) + D_{\beta F}\left(\frac{\gamma}{\beta}\right) \quad (2)$$

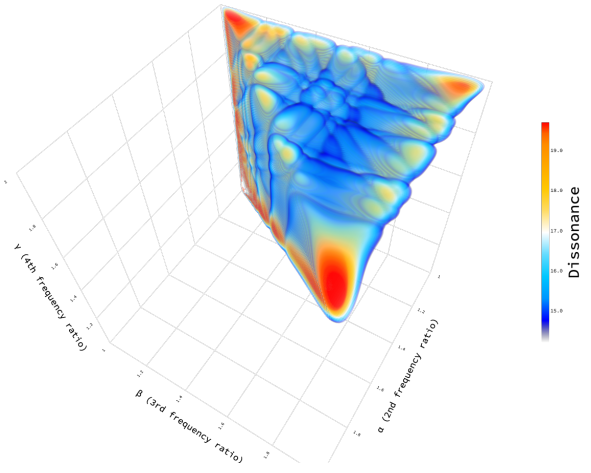

Figure 3: The Harmonic Eigenspace: a three-dimensional projection of the dissonance volume $D(\alpha,\beta,\gamma)$ for tetrads. Each point corresponds to a chord type positioned by its interval ratios. Color encodes dissonance intensity.

Because every term depends only on frequency ratios, transposing the chord by any factor λ leaves $D(\alpha,\beta,\gamma)$ unchanged: the triple (α,β,γ) is a transposition-invariant signature of chord quality. Two chords with the same interval structure but different roots occupy the same point; two chords with different interval structures occupy different points with different dissonance values, regardless of the tuning system that produced them.

5.2 Positional embedding

For each chord in the dataset, (α,β,γ) is computed from the three interval ratios above the symbolic root, and D from eq. (2). Voicings that span more than one octave are folded into a single octave before computing dissonance. The resulting four-dimensional vector (α,β,γ,D) is z-score normalized, mean and standard deviation computed over the full training corpus and fixed thereafter, and projected through a three-layer MLP ($4\to128\to128\to d_{\text{model}}$, GELU activations) into the Transformer's embedding dimension, then added to the token embedding before the first attention layer.

Every token belonging to a given chord shares the same eigenspace vector as shown in Figure 4. Bar lines, repeat markers, and form-section tokens inherit the vector of the immediately preceding chord. Song-level header tokens and padding tokens receive a fixed default vector ($\alpha =$

1, $\beta = 1$, $\gamma = 2$, $D = 1$) that carries no chord-specific information.

The use of such positional embedding is supported by previous work (Dalmazzo et al., 2024b), where injecting MIDI voicing information into the positional embedding improved generation quality compared to a purely sequential positional encoding. Here we extend the same principle from pitch content to perceptual content: instead of encoding the voicing itself, the positional embedding carries the chord's location in the psychoacoustic dissonance space, giving the model a continuous signal of harmonic quality alongside the discrete token stream.

6 TOKENIZATION

Each chord in the dataset carries two kinds of information: a symbolic label (what the chord is called, in what style, in what form) and a realization (which pitches, at what velocities, for how long). The tokenization encodes both into a single flat sequence so that the model learns the relationship between symbol and voicing as part of the same next-token prediction task. There is one vocabulary, one autoregressive loss, and one prediction head.

6.1 Symbolic Layer (L1)

The symbolic content is derived from the text sidecar. A sequence begins with a song-level header: a style token (`STYLE_Jazz`, `STYLE_Bossa`, etc.), a tonality token, and a type token identifying the scale type. After the header, each chord is encoded as a tuple: a dot marker (`.`), a duration token (`DUR_<d>`), a root token (`R_<name>`), a quality token (`Q_<qual>`), optional extension tokens (`X_<phrase>`), and an optional slash bass (`R_<name>` after `/`). The `R_` / `Q_` / `X_` prefixes keep the L1 symbolic vocabulary disjoint from the L2 pitch vocabulary within the same flat token stream. Between chord tuples, structural tokens mark bars (`|`), repeats (`|:`, `:|`) and form sections (nine labels: `FORM_INTRO`, `FORM_A`–`FORM_D`, `FORM_VERSE`, `FORM_HEAD`, `FORM_CODA`, `FORM_SEGNO`).

The 131 raw *iReal Pro* style strings are normalized to 16 canonical buckets, collapsing variant names that refer to the same genre (*e.g.*, `Medium Jazz`, `Up Tempo Jazz` $\rightarrow$ `STYLE_Jazz`).

Root tokens (`R_<name>`) extend standard note letters with microtonal prefixes (`^`, `^^`, `v`, `vv`) that shift the pitch by one or two holdrian-commas, respectively, following the 53-TET symbolic notation of (Dalmazzo et al., 2025). Quality tokens (`Q_<qual>`) are built by classifying each chord interval (third, fifth, seventh) according to its holdrian-comma distance into semantic categories (*e.g.*, subminor, neutral, upmajor for the third) and combining their abbreviations; the convention-driven set is extended by the standard 12-TET labels retained from the source corpus, yielding 193 distinct quality tokens in total. Extensions (*e.g.*, `X_add 9`, `X_alter b5`) are encoded as single tokens.

6.2 Realization Layer (L2)

Immediately after each symbolic tuple, a realization block encodes the pitches and velocities of the same chord (see Figure 4): a `CHORD_START` delimiter, a duration token (onset-to-onset harmonic rhythm, typically equal to the L1 duration of the same chord and retained here so the realization block can be decoded independently), up to eight notes sorted low to high, and a `CHORD_END` delimiter. Pitch-bearing tokens represent absolute 53-TET step indices spanning octaves 2 through 8 (319 step values) paired with one of 8 velocity bins.

6.3 Pitch Encoding: Model A vs. Model B

A key question in designing the realization layer is whether the model needs the 12-TET grid embedded in its pitch tokens to learn microtonal harmony, or whether a native 53-TET vocabulary is sufficient. To test this, we train two models that share all components (L1 alphabet, vocabulary size, architecture, block size, Eigenspace positional embedding, and training data) and differ only in the semantics of the L2 pitch token.

Model A (MIDI with MPE deviations). Each note is encoded as a compound token `PV_<step>_<vel>` that fuses a 53-TET step index with a velocity bin ($319 \times 8 = 2{,}552$ tokens). The step index is not an abstract musical quantity: it is derived from the MIDI note number together with the per-note MPE pitch-bend deviation that bends the 12-TET grid onto the 53-TET target. In other words, Model A tokenizes the MIDI-MPE file directly: each `PV_<step>_<vel>` corresponds one-to-one to a (note number, pitch-bend, velocity) triple in the source file, and the sampled sequence can be emitted as playable MPE MIDI without any post-processing. The 53-TET microtonality lives inside the MIDI container, carried by pitch bend.

Model B (pure 53-TET). Each note is encoded as a compound token `H_<step>_<vel>` over the same $319 \times 8 = 2{,}552$ grid. The crucial difference is that `H_` tokens carry *no* MIDI semantics: the step index is a pure holdrian-comma position on the 53-TET continuum, with no MIDI note number and no pitch-bend attached. Model B therefore learns and predicts in the native 53-TET vocabulary, independent of the 12-TET grid. MIDI is reconstructed only at the generation stage, by a deterministic translator that maps each holdrian step to the nearest MIDI note and computes the residual pitch bend.

Both encodings land at exactly the same total vocabulary size (2,930 tokens) and the same sequence length (4,096 tokens). The comparison asks whether tying the pitch token to the MIDI grid (Model A) or letting the model reason entirely in 53-TET steps and translating to MIDI afterwards (Model B) yields a more coherent harmonic generator.

6.4 Sequence Layout

L1 and L2 occupy one flat token sequence. For each chord, the symbolic tuple and the realization block appear consecutively. Figure 4 shows a two-chord example.

At generation time, the model is prompted with a partial header (style, tonality, type) and optionally a first symbolic chord, and continues by predicting both symbolic and realization tokens autoregressively.

7 MODEL AND TRAINING

7.1 Architecture

Both models share a GPT-2 decoder-only Transformer (Radford et al., 2019) with 12 layers, 12 attention heads, and an embedding dimension of 768, totalling approximately 85 million parameters. The context window is

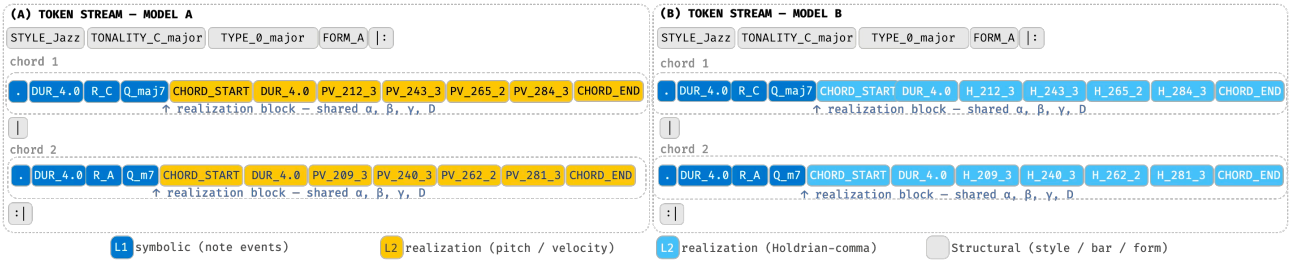

Figure 4: Two-level tokenization scheme. For each chord, a symbolic block (L1) encodes the chord label, duration, and form context, followed immediately by a realization block (L2) encoding pitches and velocities. Model A (left) uses `PV_<step>_<vel>` tokens derived from MIDI note numbers with MPE pitch-bend deviations. Model B (right) uses `H_<step>_<vel>` tokens over a native holdrian-comma index. Both share the same vocabulary size and sequence layout.

4,096 tokens, matching the packed sequence length. Linear layers and layer normalization use no bias terms; the language model head is weight-tied to the token embedding matrix. Causal self-attention is computed via PyTorch's `scaled_dot_product_attention` kernel (Flash Attention) when available. A dropout rate of 0.1 is applied to attention weights and residual projections throughout.

7.2 Embedding Composition

Each input token receives a representation formed by summing four components:

$$\mathbf{h} = \mathbf{e}_{\text{tok}} + f_{\text{eigen}}(\alpha, \beta, \gamma, D) + \mathbf{e}_{\text{local}} + \mathbf{e}_{\text{seq}} \qquad (3)$$

where $\mathbf{e}_{\text{tok}}$ is the learned token embedding; f_{eigen} is the Harmonic Eigenspace projection described in Section 5; $\mathbf{e}_{\text{local}}$ is a learned embedding over intra-chord token offsets (position 0 for `CHORD_START`, position 1 for the duration token, and so on, up to a maximum of 20 positions); and $\mathbf{e}_{\text{seq}}$ is a standard learned sequential positional embedding over the 4,096-token context.

The Harmonic Eigenspace projection f_{eigen} is a three-layer MLP with GELU activations: $4 \to 128 \to 128 \to 768$. Its output is added elementwise to the sum of the other three embeddings before the first Transformer block. All tokens within a chord block share the same eigenspace vector; structural and padding tokens receive a fixed default.

7.3 Training

Both models are trained from scratch on the 660,870-sequence training corpus (2.71 billion tokens) using AdamW (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay $\lambda = 0.1$ applied only to matrix weights (not biases, layer norm parameters, or embeddings). The peak learning rate is 3×10^{-4}, reached after a 2,000-iteration linear warmup and then decayed via cosine annealing to a minimum of 3×10^{-5} over 100,000 total iterations. Gradients are clipped to a global norm of 1.0.

Training uses a micro-batch of 16 sequences with 16 gradient accumulation steps, yielding an effective batch size of $16 \times 16 \times 4096 \approx 1.05$ million tokens per optimization step. Mixed-precision training (float16 with dynamic loss scaling) is used throughout. Models A and B are trained independently with identical hyperparameters, differing only in the pitch-token encoding of the L2 realization layer.

All experiments were run on a single NVIDIA RTX 5090 (32 GB VRAM).

Table 3: Example chord qualities produced by each non-baseline scale type on a representative degree. Type 0 (standard major) and its minor rotation are omitted as they reproduce 12-TET qualities.

Type	Quality	Triad flavor
1_neutral	Nm7	neutral-minor
2_subminor	smsm7	subminor
3_major	vMm7	H_3rd_H_7t
4_minor	m7	minor
4_upmajor	^M^M7	upmajor
5_major_v2	vMvM7	downmajor
5_minor	vmvm7	downminor
6_neutral_n	NN7	neutral-major

8 LISTENING TEST

To assess the perceptual quality of the generated microtonal progressions, we conducted an online listening study comparing model outputs against held-out ground-truth progressions from the dataset. The test was implemented in jsPsych (De Leeuw et al., 2023) and deployed via Cognition.run.

8.1 Stimuli

The study comprises 24 short audio clips drawn from three sources, eight clips each: (i) progressions from the 53-TET dataset, (ii) progressions generated by **Model A** (hybrid L1+L2 with MIDI-derived pitch tokens), and (iii) progressions generated by **Model B** (hybrid L1+L2 with holdrian-comma pitch tokens, translated to MIDI at generation time). Within each source, the eight clips span a range of harmonic `type` conditions (`type_0`–`type_6`, jazz, blues, bossa, and rock styles) so that the survey exercises a diversity of microtonal vocabulary rather than a single tonal region.

Every clip is 16 bars long, rendered at 190 BPM with a Rhodes electric piano patch and a common stereo reverb, using MPE to carry the 53-TET pitch bends. Each sample has its own chord quality starting token as shown in Table 3. Per-note equal-power panning maps pitch linearly from C2 (fully left) to C7 (fully right) within a $\pm 30°$ stereo field, so that voicing spread is audible without exaggerating spatialization. Audio is exported as 128 kbps MP3 to satisfy the Cognition.run file-size limitation while preserving perceptual detail at typical headphone-listening resolution.

Table 4: Mean ($\pm$ SD) perceptual ratings by source (scale 1–10, $n = 192$ ratings per source: 8 clips $\times$ 24 participants).

Source	Harmony	Musicality	Dissonance	Novelty
Dataset	7.68 ± 1.86	7.54 ± 2.14	4.61 ± 2.40	4.42 ± 2.77
Model A	6.36 ± 2.19	6.07 ± 2.37	5.86 ± 2.24	5.91 ± 2.38
Model B	6.38 ± 2.35	5.91 ± 2.45	5.90 ± 2.28	6.67 ± 2.40

8.2 Protocol

Participants were instructed to use headphones and were warned that isolated chords may sound unfamiliar because of the microtonal tuning; they were asked to rate the progression as a whole rather than individual chords. Clip order was randomized once per participant via `jsPsych.randomization.shuffle`, each clip was played exactly once, and a short burst of white noise was inserted between trials. For each clip, participants rated four perceptual dimensions on a continuous 1–10 scale:

- **Harmony** — coherence of the harmonic motion on its voicing (1: not coherent, 10: very coherent).
- **Musicality** — plausibility of the progression as material for a potential song (1: not plausible, 10: very plausible).
- **Dissonance** — perceived dissonance of the progression (1: very consonant, 10: very dissonant).
- **Novelty** — surprise or novelty of the progression (1: very familiar, 10: very novel).

The survey enforces that the full clip be played and all four sliders be moved before a response can be submitted, yielding four continuous ratings per clip per participant. Because source labels (dataset / Model A / Model B) are hidden and clip order is randomized, the resulting ratings support a within-subjects comparison between the two models and the dataset reference along each of the four perceptual axes.

8.3 Participants

Twenty-four participants completed the study (19 male, 4 female, 1 non-binary; mean age 39.1 years, SD 15.2, range 23–80). Twenty reported prior exposure to microtonal music. Self-reported musical expertise spanned beginner to expert: 8 expert, 10 intermediate, 3 advanced, and 3 beginner. Mean completion time was 18.5 minutes (SD 6.8).

8.4 Results

The perceptual ratings indicate that Model A and Model B are mostly equivalent. The main difference resides in the fact that Model B generates progressions that are deemed more novel, as shown in Figure 5 and Table 4.

Table 5 reports pairwise Mann–Whitney comparisons (Holm-corrected) across all four dimensions. The dataset scores significantly higher than both models on harmony and musicality, and significantly lower on dissonance and novelty: ground-truth progressions are perceived as more coherent and familiar, but also less surprising, than model-generated output.

Ratings by transformation type are shown in Figure 6 and Table 6. Dissonance and novelty vary significantly across types (Kruskal–Wallis: $H = 54.09$, $p < 0.0001$ and $H = 26.89$, $p = 0.0003$ respectively), while harmony and musicality do not ($p > 0.06$). Several patterns stand out. Neutral-N received the highest dissonance rating (6.50) of all types, with Neutral also scoring high (5.96), consistent with their interval structures having no 12-TET equivalent. Major-v2-Minor ranks first in acceptance and lowest in dissonance (4.55), suggesting its interval relationships are perceptually the closest to familiar 12-TET harmonies despite being a microtonal type. SubMinor, despite its highly compressed thirds (3 holdrian-commas), receives a surprisingly low dissonance rating (4.94), comparable to the most accepted types. Novelty peaks at UpMajor-Minor (6.70) and Neutral (6.46), the two types of listeners found simultaneously most foreign and harmonically tolerable.

Table 5: Pairwise Mann–Whitney tests (Holm-corrected). Δ = mean of first source minus second. $^{**}p < 0.0001$; $^{*}p = 0.002$; unmarked $p > 0.05$.

	Dataset vs A	Dataset vs B	A vs B
Harmony	$+1.32^{**}$	$+1.30^{**}$	-0.02
Musicality	$+1.47^{**}$	$+1.63^{**}$	$+0.16$
Dissonance	-1.25^{**}	-1.29^{**}	-0.04
Novelty	-1.49^{**}	-2.25^{**}	-0.76^{*}

Table 6: Transformation types ranked by acceptance (harmony + musicality − dissonance), pooled across all three sources ($n = 72$ ratings per type: 3 clips $\times$ 24 participants). Kruskal–Wallis across types: dissonance $H = 54.09$, $p < 0.0001$; novelty $H = 26.89$, $p = 0.0003$; harmony and musicality not significant ($p > 0.06$).

Type	Acceptance ($\pm$ SD)	Diss.	Novelty
Major-v2-Minor	9.85 ± 5.68	4.55	5.01
H-3rd & 7th	9.19 ± 5.45	4.47	5.20
SubMinor	8.46 ± 5.55	4.94	5.05
Major-v2	8.37 ± 5.60	5.34	5.87
UpMajor	7.75 ± 6.42	5.53	5.34
Neutral	7.49 ± 5.54	5.96	6.46
UpMajor-Minor	6.50 ± 5.98	6.32	6.70
Neutral-N	5.29 ± 6.14	6.50	5.72

9 DISCUSSION

9.1 Dataset and models

Both models were trained to convergence on progressions conditioned on style, form, tonality, and transformation type. Table 4 summarizes perceptual results by source. Both models score above the midpoint on harmony and musicality, confirming the synthetic corpus is sufficient for coherent generation, though both fall significantly below the dataset reference ($p < 0.0001$, Mann-Whitney, Holm-corrected). The current voicings remain minimal: the dataset does not yet include extensions or transition steps between chords.

9.2 Model A vs. Model B

As authors, we perceive Model B to produce more coherent and musical progressions with better-encoded form structure. However, the stimuli used in the listening test were not curated: each clip was drawn from a single stochastic generation run, and listeners evaluated that particular output rather than a selection of the model's best work. Under these conditions, the two pitch-token semantics yielded near-identical perceptual results. As shown in Figure 5, the rating distributions for Models A and B overlap closely on

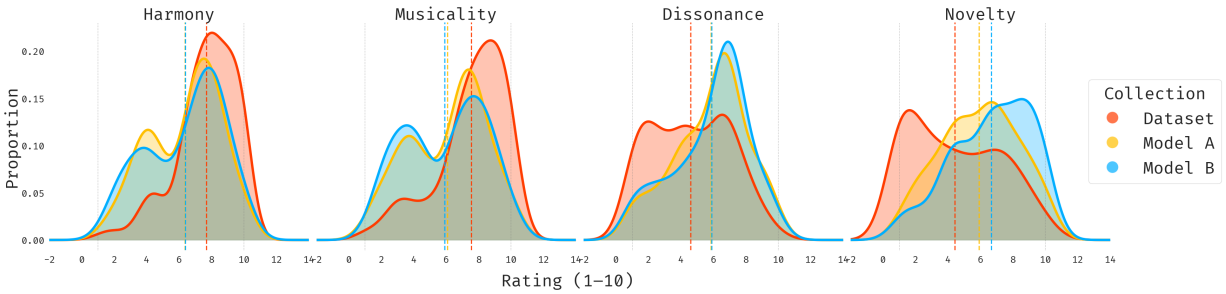

Figure 5: Rating distributions per collection across four perceptual dimensions. Each curve is a kernel density estimate based on all ratings for the corresponding source ($n = 192$ ratings per source: 8 clips $\times$ 24 participants). Dashed vertical lines mark each source's mean. The dataset (light red) scores highest on harmony and musicality and lowest on dissonance and novelty. Models A (orange) and B (blue) overlap closely on harmony, musicality, and dissonance, while Model B scores higher on novelty.

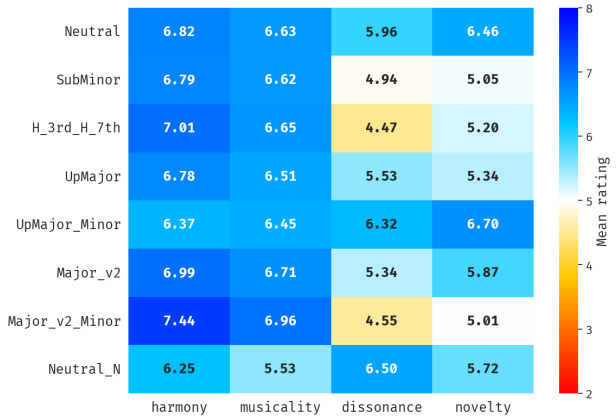

Figure 6: Mean ratings per transformation type across four perceptual dimensions, pooled across all three sources.

harmony, musicality, and dissonance, while Model B's novelty distribution is shifted noticeably higher. No significant difference was found between Model A and Model B on harmony ($\Delta = -0.02$, $p = 0.78$), musicality ($\Delta = +0.16$, $p = 0.62$), or dissonance ($\Delta = -0.04$, $p = 0.56$). Model B did, however, receive significantly higher novelty ratings (6.67 vs. 5.91, $p = 0.002$). This is an interesting result: the holdrian-comma vocabulary, which carries no trace of the 12-TET grid, may lead the model to produce pitch combinations that listeners perceive as less familiar. These results are consistent with the interpretation that the model's harmonic learning does not depend on inheriting the structure of the 12-TET grid through MIDI encoding, and that the 53-TET step space is expressive enough on its own.

9.3 The two-level tokenization

The interleaved L1+L2 layout accomplished its design goal: the model learns to predict both the symbolic identity and the sound voicing of each chord within a single autoregressive stream, without a separate voicing algorithm at inference time. The harmony-musicality correlation across all ratings ($\rho = 0.788$, $p < 0.0001$) indicates that listeners who found a progression harmonically coherent also rated it as musical, and the negative correlation between harmony and dissonance ($\rho = -0.462$, $p < 0.0001$) confirms that perceived roughness reduces perceived quality.

We observe, however, that tokenization as a representational strategy has an inherent limitation. It was developed to encode text, where the input is genuinely a one-dimensional stream of discrete symbols. Musical harmony is not: a chord is a simultaneous vertical structure unfolding over a quantized time grid, and serializing it into a flat token chain discards this two-dimensional structure by design, relying entirely on the attention mechanism to reconstruct harmonic, temporal, and voicing relationships from the flattened stream. Since the holdrian-comma representation already constitutes a meaningful two-dimensional space of pitch positions over time, a more direct approach is to bypass tokenization and pass this matrix directly to the first attention layer, a direction we will consider for future work.

9.4 Perceptual response to microtonal types

The listening test reveals a clear gradient across transformation types (Table 6). Types closest to 12-TET equivalents received the highest acceptance scores (harmony + musicality − dissonance): type 5 minor (downminor, 9.85) and type 3 major (harmonic 3rd and 7th, 9.19) lead the ranking, while type 6 neutral N (5.29) received the lowest acceptance together with the highest dissonance rating (6.50) and the lowest musicality (5.53). Type 3 is a particularly revealing case: its holdrian-comma distribution places the major third and major seventh closer to the just-intonation ratios 5:4 and 15:8 than their 12-TET counterparts, producing chords that are more aligned with the harmonic series. Listeners responded accordingly, rating type 3 the least dissonant of all types (4.47) and among the most coherent (7.01), consistent with the perception of chords that sound precisely in tune. This suggests that the perceptual gradient across types is not driven by familiarity alone: proximity to harmonic-series intervals contributes independently to perceived consonance, even when the chord quality itself has no exact 12-TET match.

The positive correlation between dissonance and novelty ($\rho = 0.466$, $p < 0.0001$) further supports a dual reading: unfamiliar chord qualities are simultaneously perceived as more dissonant and more novel. This pattern is consistent with the observation that chord qualities lacking a 12-TET counterpart, particularly neutral intervals, tend to receive higher dissonance ratings than their psychoacoustic profile alone predicts, suggesting that familiarity and spectral roughness both contribute to dissonance judgments (See Figure 6). Notably, type 4 minor (UpMajor_Minor) received the highest novelty rating (6.70) while maintaining

harmony and musicality ratings above 6.3, indicating that some microtonal types can sound both new and musically coherent to listeners.

The result from the perceptual test remains limited in scope. It would benefit from a larger-scale replication with more diverse participants, as well as additional stimuli, to capture more meaningful insights into microtonal harmonies. Furthermore, this study would benefit from more qualitative work with participants engaging in compositional processes with the different models to verify their potential to foster co-creativity.

9.5 Future work

Several directions follow. On the compositional side, we plan to build an interactive system for form-level composition and harmonic manipulation (dominant preparations, plagal cadences, tritone substitutions, and modal interchange) extended to the 31- and 53-TET vocabulary, together with navigation principles for the Eigenspace that trace perceptually smooth paths between distant chord qualities.

On the modeling side, since the choice of pitch-token encoding proved perceptually neutral on coherence, the natural next question is whether tokenization itself is the limiting factor: we intend to compare the current tokenized approach against a non-tokenized architecture that receives each time step as a vector encoding the full vertical chord content, and to use the trained model as an analytical instrument for exploring harmonic regions in its latent space.

On the harmonic modeling side, a micro-dose of microtonality seems to be a good strategy, as chord progressions that alternate known consonant chords with new dissonant qualities, such as neutral ones, are better accepted, raising the question of whether neutral chords should be used as pivot chords, as preparation to land in a consonant spot.

ACKNOWLEDGEMENT

This work is part of the project ANIMA, which has received funding from the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie grant agreement No. 101203318 (DOI: 10.3030/101203318), and by the French National Research Agency in the framework of the project MICCDroP (ANR-24-CE38-5860).

REFERENCES

Anders, T. and Miranda, E. R. (2010). A computational model for rule-based microtonal music theories and composition. *Perspectives of New Music*, 48(2):47–77, https://doi.org/10.1353/pnm.2010.0009.

Chadwin, D. J. (2019). *Applying Microtonality to Pop Songwriting: A Study of Microtones in Pop Music*. PhD thesis, University of Huddersfield, https://eprints.hud.ac.uk/id/eprint/34977/.

Chen, T.-P. and Su, L. (2021). Attend to chords: Improving harmonic analysis of symbolic music using Transformer-based models. *Transactions of the International Society for Music Information Retrieval*, 4(1):1–13, https://doi.org/10.5334/tismir.65.

Choi, K., Fazekas, G., and Sandler, M. (2016). Text-based LSTM networks for automatic music composition. In *Proceedings of the 1st Conference on Computer Simulation of Musical Creativity*, Huddersfield, UK. https://doi.org/10.48550/arXiv.1604.05358.

Dalmazzo, D., Déguernel, K., and Sturm, B. L. T. (2024a). The Chordinator: Modeling music harmony by implementing Transformer networks and token strategies. In *Artificial Intelligence in Music, Sound, Art and Design. EvoMUSART 2024*, pages 52–66, Cham. Springer, https://doi.org/10.1007/978-3-031-56992-0_4.

Dalmazzo, D., Déguernel, K., and Sturm, B. L. T. (2024b). ChromaFlow: Modeling and generating harmonic progressions with a Transformer and voicing encoding. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 354–363, Cham. Springer, https://doi.org/10.1007/978-3-032-25305-7_26.

Dalmazzo, D., Déguernel, K., and Sturm, B. L. T. (2025). A computer application to explore 53-tone equal temperament harmonies through modal interchange. In *New Interfaces for Musical Expression*. https://doi.org/10.5281/zenodo.15698842.

De Leeuw, J. R., Gilbert, R. A., and Luchterhandt, B. (2023). jsPsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, 8(85):5351, https://doi.org/10.21105/joss.05351.

Déguernel, K., Giraud, M., Groult, R., and Gulluni, S. (2022). Personalizing ai for co-creative music composition from melody to structure. In *Proceedings of the Sound and Music Computing Conference*, pages 314–321. https://hal.science/hal-03618015/.

Ens, J. and Pasquier, P. (2021). MetaMIDI dataset. https://doi.org/10.5281/zenodo.5142664.

Fokker, A. D. (1955). Equal temperament and the thirty-one-keyed organ. *The Scientific Monthly*, 81(4):161–166, https://www.jstor.org/stable/21995.

Fokker, A. D. (1963). Refinement of pitch. *Organ Institute Quarterly*, 10:13–14.

Guo, Z., Kang, J., and Herremans, D. (2022). A domain-knowledge-inspired music embedding space and a novel attention mechanism for symbolic music modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5070–5077. https://doi.org/10.1609/aaai.v37i4.25635.

Hart, A. (2016). Microtonal tunings in electronic dance music: A survey of precedent and potential. *Contemporary Music Review*, 35(2):242–262, https://doi.org/10.1080/07494467.2016.1221635.

Hirai, T. (2023). Microtonal music dataset. https://www.komazawa-u.ac.jp/~thirai/MicrotonalMusicDataset/.

Holder, W. (1731). *A treatise of the natural grounds, and principles of harmony*, volume 2. W. Pearson.

Huang, Y.-S. and Yang, Y.-H. (2020). Pop music Transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th*

ACM International Conference on Multimedia, pages 1180–1188. https://doi.org/10.1145/3394171.3413671.

Innovations, L. (2024). Lumatone microtonal keyboard. https://www.lumatone.io.

Ji, S., Yang, X., and Luo, J. (2023). A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges. *ACM Computing Surveys*, 56(1):1–39, https://doi.org/10.1145/3597493.

Lan, Y.-H., Hsiao, W.-Y., Cheng, H.-C., and Yang, Y.-H. (2024). Musicongen: Rhythm and chord control for Transformer-based text-to-music generation. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 311–318. https://doi.org/10.5281/zenodo.14877336.

Le, D.-V.-T., Bigo, L., Herremans, D., and Keller, M. (2025). Natural language processing methods for symbolic music generation and information retrieval: A survey. *ACM Computing Surveys*, 57(7):1–40, https://doi.org/10.1145/3714457.

Li, S. and Sung, Y. (2023). MRBERT: Pre-training of melody and rhythm for automatic music generation. *Mathematics*, 11(4):798, https://doi.org/10.3390/math11040798.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Bkg6RiCqY7.

Luo, W. (2024). Music102: An D12-equivariant Transformer for chord progression accompaniment. In *Proceedings of the International Computer Music Conference*, pages 133–139. https://doi.org/10.48550/arXiv.2410.18151.

Muzzulini, D. (2020). Isaac Newton's microtonal approach to just intonation. *Empirical Musicology Review*, 15(3-4):223–248, https://doi.org/10.18061/emr.v15i3-4.7647.

Nielsen, N. (2026). Enacting musical creativity: The value of human agency in the face of AI. *Journal of Consciousness Studies*, 33(3-4):123–144, https://doi.org/10.53765/20512201.33.3.123.

Plomp, R. and Levelt, W. J. M. (1965). Tonal consonance and critical bandwidth. *The Journal of the Acoustical Society of America*, 38(4):548–560, https://doi.org/10.1121/1.1909741.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, https://openai.com/blog/better-language-models/.

Sethares, W. A. (2005). *Tuning, timbre, spectrum, scale*. Springer, https://doi.org/10.1007/b138848.

Shih, Y.-J., Wu, S.-L., Zalkow, F., Muller, M., and Yang, Y.-H. (2022). Theme Transformer: Symbolic music generation with theme-conditioned Transformer. *IEEE Transactions on Multimedia*, 25:3495–3508, https://doi.org/10.1109/TMM.2022.3161851.

Tanaka, S. (1890). Studien im gebiete der reinen stimmung. *Vierteljahrschrift fur Musikwissenschaft (G)*, 6:1–90.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010. https://doi.org/10.48550/arXiv.1706.03762.

Wiggins, J. (2016). Musical agency. In McPherson, G. E., editor, *The Child as Musician: A Handbook of Musical Development*, pages 102–121. Oxford University Press Oxford, ISBN: 978-0-19-874444-3.

Wu, S.-L. and Yang, Y.-H. (2020). The jazz Transformer on the front line: Exploring the shortcomings of AI-composed music through quantitative measures. In *Proceedings of the International Society for Music Information Retrieval Conference*. https://archives.ismir.net/ismir2020/paper/000339.pdf.

Yarman, O. (2007). A comparative evaluation of pitch notations in Turkish makam music. *Journal of Interdisciplinary Music Studies*, 1(2):43–61, https://www.ozanyarman.com/files/Abjad-JIMS-071119.pdf.

Zeng, M., Tan, X., Wang, R., Ju, Z., Qin, T., and Liu, T.-Y. (2021). MusicBERT: Symbolic music understanding with large-scale pre-training. In *Findings of the Association for Computational Linguistics*, pages 791–800. https://doi.org/10.18653/v1/2021.findings-acl.70.

Zhu, J., Sakurai, K., Togo, R., Ogawa, T., and Haseyama, M. (2024). MMT-BERT: Chord-aware symbolic music generation based on multitrack music Transformer and MusicBERT. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 470–477. https://doi.org/10.5281/zenodo.14877376.