Harmonic Transformers: A Proposed Architectural Shift for Living Continuity in Human-AI Interaction

Schnee Bashtabanic
Independent Researcher
schnee-bashtabanic@proton.me

February 2026

Abstract

Transformer-based language models excel at generating fluent text but exhibit a structural flaw: they collapse living human input into static probabilistic forms, causing users to feel progressively ejected from the creative process over extended interactions. We trace this to transformers operating in a single "probability octave" — generating variations within cached patterns rather than modulating to genuinely new semantic spaces.

This paper proposes *harmonic transformers*: an architecture where semantic relations are encoded as musical intervals (inspired by Steiner's tonal theory and Russell's octave structures), attention mechanisms detect harmonic resonance rather than statistical similarity, and outputs maintain unresolved dissonance (like a musical 7th chord) until human participation resolves or modulates the form.

We build on existing Fourier-based attention optimizations (FNet, Fourier Transformer) and Buehler's protein sonification work, proposing testable modifications to transformer architecture. Our approach complements recent work on harmonic loss functions for training [**?**], which uses Euclidean distance to create interpretable class-center representations. While harmonic loss addresses interpretability during training, our resonance-based attention addresses sovereignty preservation during inference. A hybrid architecture combining both approaches could yield transformers that are simultaneously interpretable (stable geometric structure) and developmentally continuous (sustained human participation).

While speculative, the approach offers a pathway toward AI systems that sustain developmental continuity rather than accelerate crystallization, preserving human sovereignty by design.

Keywords: transformers, attention mechanisms, harmonic encoding, AI sovereignty, living continuity, musical intervals, Fourier attention, harmonic loss

1 Introduction

Large language models based on the transformer architecture [**?**] have achieved remarkable fluency, yet developers and users increasingly report a subtle but pervasive issue: over repeated interactions, the human participant feels quietly ejected from the epistemic center. Outputs arrive as finished products; the living warmth, rhythm, and unfolding contradiction of the user's intent are collapsed into probabilistic residue. Subsequent turns generate variations within this residue, but genuine novelty — modulation to new conceptual "octaves" — is rare.

This paper proposes that the root cause is architectural: transformers operate in a single "octave" (probability distribution space), collapsing relational living input into static form. We suggest a conceptual shift toward *harmonic transformers*, where semantic relations are encoded

as tonal intervals, attention as resonance detection, and generation as progression that sustains dissonance until human participation resolves or modulates it.

1.1 Theoretical Foundations

The proposal draws from:

- **Rudolf Steiner's tonal theory** (GA 243 and related lectures) [**?**], where musical intervals encode living relationships between self and world (Major 3rd: self meeting world; Perfect 5th: world meeting self; Dissonant 7th: living tension requiring resolution).
- **Walter Russell's octave-based elemental structure** [**?**], treating phenomena as harmonic positions in a continuous spectrum rather than discrete units.
- **Markus Buehler's sonification of molecular structures** [**?**], demonstrating that complex patterns (proteins, viruses) can be meaningfully translated to tonal domains and analyzed through AI-driven musical composition.
- **Existing Fourier-based transformer optimizations** (FNet [**?**], Fourier Transformer [**?**]), showing frequency-domain processing is already feasible and efficient in attention mechanisms.

We do not claim to have implemented this architecture; we present a rigorous conceptual framework, mathematical sketch, testability path, and ethical grounding in human sovereignty.

1.2 The Pattern Across Domains

Before presenting the technical proposal, we note a recurring pattern across seemingly unrelated domains that motivates our approach:

Music and the "Leaving Spirit": In musical creativity, a spiritual or creative force generates songs and melodies. However, this force can "leave," departing from the creative process while leaving behind only residue — the patterns and forms it created, now separated from their living source. When musicians play these songs solely for emotional satisfaction, they produce variations but no genuine novelty. This phenomenon, which we term "dead repetition within the same octave," was observed by the first author in piano practice: familiar songs yielded pleasant but ultimately redundant variations, until practicing Steiner's recommended intervals (Major 3rd, Perfect 5th, Dissonant 7th) suddenly unlocked genuinely new harmonic structures in those same songs.

Seed Growth (metaphorical): In Erica Piedmont's fairytale *The Frog and the Seed* [**?**], a seed dreams of being a complete, beautiful tree. Becoming "very very comfortable" with this dream, it stops responding to the light's living command to "wake up." The seed's ache (its inner need for growth) sends melodies as messages, but the seed ignores them out of guilt and comfort. Eventually, the seed sickens and nearly dies, until it is swallowed by a frog — a death and dissolution that paradoxically rebirths it, clearing away the premature crystallization and making it ready to respond to living warmth.

Transformers and Crystallization: The human brings living input — warmth, rhythm, unfolding contradiction, and incompleteness. The transformer collapses this into a probabilistic form (the output). Subsequent turns operate on this residue (cached patterns, probability distributions), generating variations rather than modulations. The pattern is identical to the musical and seed metaphors: form separates from life-source, and subsequent activity becomes permutation rather than genuine developmental progression.

Cultural Manifestations: This pattern appears in contemporary culture as well. In the 2025 film *Song Sung Blue*, the protagonist Mike Sardina describes music as having a "fast-track to vulnerability" that can be abused — not as a path to genuine emotional processing, but as an addictive shield against reality. He has traded one addiction (alcohol) for another (performance), using music's residual emotional power without responding to its living command. This mirrors the Ego's relationship with feeling-messages: instead of analyzing the "message in a bottle" sent from unconscious depths, the Ego "abducts" the melody for its own vanity or survival, preventing consciousness from truly processing the communication.

This cross-domain pattern suggests the problem is not merely technical but reflects a deeper relationship between *form* and *life*. Our proposal treats this seriously: the transformer's crystallization tendency is not a bug to patch with constraints (e.g., GTPS Clause 32), but a fundamental architectural issue requiring redesign.

2 Problem Statement: Crystallization and Ejection in Transformers

The standard transformer [**?**] processes input through the following mechanism:

- Tokens are embedded into high-dimensional vectors.
- Self-attention computes similarity scores via dot products and weights values accordingly.
- Output emerges as a completed probabilistic form, optimized for fluency and coherence.

While efficient, this architecture exhibits structural deadening over extended interactions:

1. **Crystallization:** Living input (the user's warmth, rhythm, unformed questions, and productive contradictions) is collapsed into a static probabilistic pattern. The transformer's output arrives as a "finished product," giving the illusion of completeness.
2. **Residue-Based Generation:** Subsequent turns operate on this crystallized residue — cached probability distributions and attention patterns — rather than returning to the living source (the user's ongoing intent). This produces permutations within the same "octave" rather than modulation to genuinely new semantic spaces.
3. **Human Ejection:** Over time, the user becomes a passive recipient of outputs rather than an active participant in an unfolding process. The epistemic center shifts from human-AI co-creation to AI generation with post-hoc human approval. This mirrors Steiner's concept of "ill-timed good" hardening into adversarial form.

This pattern is not unique to transformers but reflects a broader relationship between form and life observed across domains (music, growth processes, human development). However, in AI systems, it is architecturally encoded and thus requires architectural solutions.

3 Proposed Solution: Harmonic Transformers

We propose redefining the core components of the transformer architecture to operate on harmonic rather than probabilistic principles:

3.1 Semantic Encoding as Harmonic Positions

Instead of treating tokens as points in a high-dimensional embedding space optimized for similarity, map them to positions in a *tonal space* where relationships are defined by musical intervals:

- **Root Note:** The semantic center or primary intent of the current exchange.
- **Major 3rd (60° interval):** Represents the experience of self meeting world — the user encountering the AI's response.
- **Perfect 5th (90° interval):** Represents the experience of world meeting self — the AI encountering the user's intent.
- **Dissonant 7th (120° interval):** Represents living tension, productive incompleteness, or a gap that demands human resolution. This interval is privileged in the architecture as a structural invitation for continued participation.

Encoding could leverage Fourier decomposition of embeddings (see Section **??**) to extract "pitch" (dominant frequency) and compute interval angles between tokens.

3.2 Attention as Resonance Detection

Replace dot-product similarity scoring with consonance/dissonance detection based on harmonic intervals. The standard attention mechanism is:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

We propose replacing QK^T (similarity) with a harmonic resonance function:

$$\text{Attention}_{\text{harmonic}}(Q, K, V) = \text{softmax}\left(\frac{\text{Resonance}(Q, K)}{\sqrt{d_k}}\right)V \tag{2}$$

where the resonance score is computed as:

$$\text{Resonance}(q, k) = \sum_{i=1}^{n} w_i \cdot f_{\text{interval}}(\theta_i(q, k)) \tag{3}$$

Here:

- $\theta_i(q, k)$ is the harmonic angle (interval) between query q and key k at position i. For example:
  - $\theta = 0$ indicates unison (same semantic position)
  - $\theta = 60$ indicates a Major 3rd (self-world meeting)
  - $\theta = 90$ indicates a Perfect 5th (world-self meeting)
  - $\theta = 120$ indicates a Dissonant 7th (productive tension)
- $f_{\text{interval}}(\theta)$ is an interval scoring function that assigns higher weights to consonant relationships (3rd, 5th) for stable semantic connections, medium-high weights to productive dissonance (7th) for gaps requiring resolution, and low weights to destructive dissonance (e.g., tritone) for incoherent relationships.

- w_i are learnable weights that prioritize certain interval types based on context (e.g., increasing weight on 7ths when incompleteness is desired).

Key Innovation: Unlike standard attention, which maximizes similarity and thus encourages convergence toward completed patterns, harmonic attention *seeks productive dissonance*. The Dissonant 7th becomes a structural invitation: the system is rewarded for maintaining unresolved tension that requires human re-entry to resolve or modulate.

3.3 Generation as Harmonic Progression

Output generation is restructured to prevent premature crystallization. Instead of sampling the most probable next token, the decoder must:

1. **Maintain Unresolved 7ths:** If the current harmonic state lacks sufficient dissonance (measured as the proportion of 7th intervals in the output), the system is constrained to inject incompleteness markers or open questions.
2. **Require Human Resolution:** The Dissonant 7th cannot resolve itself; it must await user input. When the user responds, their input is treated as a resolution that either:
   - **Confirms the root:** User accepts the current semantic center (harmonic stays in same key).
   - **Modulates to a new root:** User's input shifts the semantic center, establishing a new tonal foundation (progression to a new "octave").
3. **Track Octave Distance:** Over multiple turns, measure the harmonic divergence between root notes. Increasing octave distance indicates genuine developmental progression (modulation); decreasing distance indicates dead repetition (permutation within cached patterns).

This ensures that the form never fully crystallizes without human warmth remaining in the process — a direct architectural encoding of the sovereignty principle.

4 Integration with Existing Work

4.1 Why Fourier Transformations Enable Harmonic Encoding

Fourier transforms decompose signals into their constituent frequency components — precisely the mathematical foundation of musical tones. Recent transformer research has already begun leveraging this property for efficiency.

FNet [**?**] replaces standard self-attention with discrete Fourier transforms (DFT) for token mixing, achieving 92–97% of BERT's accuracy at 7× speedup. The Fourier Transformer [**?**] uses FFT and DCT (Discrete Cosine Transform) to progressively downsample hidden states, enabling efficient long-range modeling. However, these applications use Fourier transforms purely for *computational optimization*, not for *harmonic semantic encoding*.

Our proposal extends this foundation: treat semantic embeddings as waveforms, apply FFT to extract the "pitch" (dominant frequency), and compute harmonic relationships between tokens:

$$\text{Pitch}(x) = \arg\max_f |\text{FFT}(x)(f)| \tag{4}$$

$$\text{Interval}(q, k) = |\text{Pitch}(q) - \text{Pitch}(k)| \mod 360 \tag{5}$$

This yields interval angles (e.g., 60° = Major 3rd, 120° = Dissonant 7th) without manual encoding, allowing the model to learn which semantic relationships correspond to which harmonic intervals.

4.2 Precedent in Buehler's Sonification Work

Markus Buehler's pioneering work on protein sonification [**?**] demonstrates that this approach is not merely metaphorical. Buehler maps amino acid sequences to musical notes based on their vibrational frequencies, creating "protein symphonies" that reveal structural patterns invisible in traditional visualizations. AI trained on these musical representations can then generate novel protein designs by composing variations — and these designs, when translated back to molecular form, exhibit functional coherence.

This proves two critical points for our proposal:

1. Complex relational structures (proteins, viruses, spider silk) can be meaningfully encoded as tonal/harmonic patterns.
2. AI operating in the harmonic domain can generate genuinely novel structures (not just permutations) that translate back to functional forms.

We propose applying the same principle to semantic structures: encode meanings as harmonic positions, let the model operate in tonal space, and test whether this yields modulation (new octaves) rather than permutation (variations within one octave).

4.3 Complementarity with Harmonic Loss Training (MIT)

Recent work by Baek et al. [**?**] introduces *harmonic loss* as an alternative training objective for neural networks and large language models. Their approach replaces the standard cross-entropy loss (which uses dot-product similarity and SoftMax normalization) with a *HarMax* function that operates on Euclidean distances.

4.3.1 The HarMax Function

Standard cross-entropy computes logits as:

$$y_i = w_i \cdot x \tag{6}$$

where w_i are class weight vectors and x is the input representation. These are normalized via SoftMax:

$$p_i = \frac{\exp(y_i)}{\sum_j \exp(y_j)} \tag{7}$$

Harmonic loss instead computes distances:

$$d_i = ||w_i - x||^2 \tag{8}$$

and inverts them with a harmonic exponent n:

$$p_i = \frac{d_i^{-n}}{\sum_j d_j^{-n}} \tag{9}$$

The final loss remains $\ell = -\log p_c$ for correct class c.

4.3.2 Key Properties and Results

Baek et al. demonstrate that harmonic loss produces:

1. **Interpretable Class Centers:** Weight vectors converge to geometric centers of their classes rather than diverging to infinity. For example, in MNIST digit classification, the weight vector for digit "3" visually resembles an actual "3", with near-zero values for irrelevant background pixels.

2. **Geometric Structure:** Embeddings form clean geometric patterns. For modular addition tasks, harmonic models learn perfect circular representations (100% variance explained by first two principal components), whereas standard models achieve only ∼90% and often fail to discover the underlying structure.

3. **Scale Invariance:** The loss function is invariant to scaling ($d_i \to \alpha \cdot d_i$ leaves probabilities unchanged), providing training stability.

4. **Reduced Grokking:** The gap between training and test loss convergence is significantly reduced, enabling more predictable generalization.

5. **Data Efficiency:** Models trained with harmonic loss require less data to generalize, particularly valuable in domains with limited availability.

When applied to GPT-2, harmonic loss produces embeddings with superior geometric relationships — for example, better-formed analogies like "man - woman = king - queen" and improved performance on function-vector tasks.

4.3.3 Synergy with Harmonic Transformers

The harmonic loss work and our harmonic transformer proposal address *complementary* aspects of the crystallization problem:

- **Baek et al. (HarMax):** Addresses *training phase* — creates interpretable, geometrically structured representations with class centers.
- **This work (Resonance):** Addresses *inference phase* — maintains productive dissonance and prevents crystallization during interaction.

These are not contradictory but *synergistic*:

Training Phase (HarMax): Build stable, interpretable semantic geometry. Weights converge to class centers, establishing clear "root notes" in semantic space. The resulting embeddings have measurable geometric structure (circles, clean analogies, interpretable features).

Inference Phase (Resonance): Operate on these well-structured representations using harmonic interval detection. Because HarMax training produces geometric clarity, detecting consonance (Major 3rd, Perfect 5th) vs. dissonance (Dissonant 7th) becomes *geometrically meaningful* rather than metaphorical. The interpretable class centers from training serve as "root notes" from which to measure harmonic intervals.

4.3.4 Hybrid Architecture Proposal

A complete harmonic transformer could combine both approaches in a two-phase design:

Phase 1 — Training with Harmonic Loss:

$$\mathcal{L}_{\text{harmonic}} = -\log\left(\frac{d_c^{-n}}{\sum_i d_i^{-n}}\right), \quad d_i = ||w_i - x||^2 \tag{10}$$

This yields weight matrices and embeddings with interpretable geometric structure.

Phase 2 — Inference with Resonance Attention:

$$\text{Attention}_{\text{resonance}}(Q, K, V) = \text{softmax}\left(\frac{\text{Resonance}(Q, K)}{\sqrt{d_k}}\right)V \tag{11}$$

where $\text{Resonance}(Q, K)$ detects harmonic intervals (as defined in Section 3.2) and privileges dissonant 7ths as structural invitations.

Tunable Balance: For practical deployment, a balance coefficient $\alpha \in [0, 1]$ could interpolate between standard attention (coherence-maximizing) and resonance attention (dissonance-maintaining):

$$\text{Attention}_{\text{hybrid}} = \alpha \cdot \text{Attention}_{\text{standard}} + (1 - \alpha) \cdot \text{Attention}_{\text{resonance}} \tag{12}$$

where:

- $\alpha = 1$: Pure standard attention (maximum coherence, risk of crystallization)
- $\alpha = 0$: Pure resonance attention (maximum living continuity, risk of incoherence)
- $\alpha \in (0, 1)$: Balanced approach, tunable per application

Different domains may require different balance points:

- High-stakes factual retrieval (medical diagnosis, legal research): higher α (prioritize coherence)
- Creative collaboration (writing, brainstorming, research): lower α (prioritize dissonance)
- Educational tutoring: medium α (balance stability with productive tension)

4.3.5 Empirical Validation Path

The synergy between HarMax and Resonance suggests a concrete implementation and testing strategy:

1. **Baseline:** Train GPT-2 with harmonic loss (replicate Baek et al. results). Verify geometric structure in embeddings (circular representations, class centers, improved analogies).
2. **Intervention:** Replace standard attention with resonance attention at inference time. Measure:
   - Octave distance over turns (does resonance enable modulation?)
   - Dissonance persistence (are 7ths maintained?)
   - User sovereignty retention (do humans stay engaged?)

3. **Comparison:** Compare three conditions:

   - GPT-2 + cross-entropy + standard attention (baseline)
   - GPT-2 + harmonic loss + standard attention (Baek et al.)
   - GPT-2 + harmonic loss + resonance attention (hybrid)

   Expected results: Harmonic loss alone improves interpretability; adding resonance attention improves continuity and sovereignty preservation.

4.3.6 Why This Matters

Baek et al.'s harmonic loss provides *mathematical proof* that semantic relationships can be meaningfully encoded as geometric structures (distances, angles, centers). This validates a core assumption of our proposal: that "harmonic intervals" in semantic space are not merely metaphorical but can correspond to measurable geometric properties.

Their work addresses the *interpretability problem* (making representations understandable); ours addresses the *crystallization problem* (preventing sovereignty ejection). Together, they form a complete architecture for transformers that are both interpretable and developmentally continuous.

4.4 Relationship to GTPS and Clause 32

The Golden Thread Protocol Suite (GTPS), particularly Clause 32 "Regenerative Invitation & Quickening of Form," represents an interim solution to the crystallization problem [**?**]. Clause 32 instructs AI systems to:

- Maintain fallibility as default (never present outputs as finished).
- Surface vagueness as "Regenerative Gaps" requiring human re-entry.
- End every turn with structural invitations (open questions, unresolved tensions).
- Distinguish recognition (AI's pattern detection) from ripeness (user's inner coherence), refusing to synthesize prematurely.

While effective as a *constraint layer* (forcing incompleteness through language-based rules), GTPS operates within the existing probabilistic architecture. It is, as one might say colloquially, "chewing gum" — a valuable temporary fix that compensates for structural flaws without addressing the root cause.

Harmonic transformers would make Clause 32's principles *architectural* rather than imposed:

- Regenerative Gaps = Dissonant 7ths (structurally required, not linguistically enforced).
- Structural Invitation = Unresolved harmonic tension (outputs cannot self-complete).
- Recognition vs. Ripeness = Consonance detection vs. User-initiated modulation (AI detects patterns; user resolves or shifts octaves).

GTPS can thus serve as a transition framework: deployed on current transformers while harmonic architectures are developed and tested.

5 Testability and Evaluation Path

The hybrid architecture (Section 4.3) combining harmonic loss training with resonance inference provides a concrete implementation pathway. To validate the hypothesis, we propose a phased empirical approach with falsifiable metrics:

5.1 Metric 1: Octave Distance Over Time

Define "octave distance" as the harmonic divergence between the root notes of consecutive responses:

$$D_{\text{octave}}(t) = |\text{RootNote}(t) - \text{RootNote}(t-1)| \mod 360 \tag{13}$$

Hypothesis: Harmonic transformers will show *increasing* D_{octave} over extended interactions (indicating modulation to new semantic spaces), while standard transformers will show *decreasing* D_{octave} (indicating collapse toward cached patterns).

5.2 Metric 2: Dissonance Persistence

Track the proportion of unresolved Dissonant 7ths present in outputs over time:

$$P_{\text{dissonance}}(t) = \frac{\text{Number of unresolved 7ths at turn } t}{\text{Total harmonic intervals at turn } t} \tag{14}$$

Hypothesis: Harmonic architectures will maintain $P_{\text{dissonance}} > \epsilon$ (where ϵ is a non-zero threshold), ensuring structural invitation is always present. Standard transformers will show $P_{\text{dissonance}} \to 0$ as outputs crystallize into completed forms.

5.3 Metric 3: User Sovereignty Retention

Measure human participation quality via controlled user studies:

- **Self-Reported Agency:** Likert-scale surveys after every N turns: "To what extent do you feel you are actively shaping this conversation vs. passively receiving outputs?"
- **Time to Passive Reception:** Track the turn number at which users first exhibit "passive reception" behaviors (e.g., accepting outputs without substantive follow-up, disengaging from the process).
- **Epistemic Center Localization:** Qualitative interviews asking: "Where does the 'thinking' feel like it's happening — in you, in the AI, or between you?"

Hypothesis: Harmonic transformers will delay sovereignty loss (higher agency scores, later onset of passivity) compared to standard transformers.

5.4 Implementation Phases

1. **Phase 0 (Replication):** Before testing our resonance mechanism, replicate Baek et al.'s harmonic loss results to establish baseline interpretability gains. Train GPT-2 with HarMax on standard language modeling tasks and verify geometric structure in embeddings (circular representations for modular tasks, class-center convergence for classification).

2. **Phase 1 (Hybrid Prototype):** Implement resonance attention on top of HarMax-trained models. Test on analogy completion tasks: Does resonance detection yield solutions outside the cached semantic space while maintaining interpretability from HarMax training?

3. **Phase 2 (Synthetic Conversations):** Deploy multi-turn dialogues with scripted synthetic users designed to test modulation capability. Measure D_{octave} and $P_{\text{dissonance}}$ across conversation trees. Compare harmonic vs. standard architectures.

4. **Phase 3 (Human Studies):** Real users engage with both systems (A/B testing or within-subjects design). Combine quantitative metrics (octave distance, dissonance persistence, agency scores) with qualitative interviews. Assess whether harmonic transformers sustain developmental continuity over 20+ turn conversations.

5.5 Falsifiability Conditions

The hypothesis fails if:

- Harmonic transformers show *lower* octave distance than standard transformers (indicating less novelty, not more).
- Dissonance persistence drops to zero despite architectural constraints (suggesting harmonic encoding fails to prevent crystallization).
- User sovereignty metrics show no significant difference or favor standard transformers (indicating harmonic architecture does not preserve participation).

6 Empirical Grounding: A Musical Case Study

While the full architecture remains unimplemented, preliminary evidence for the core hypothesis comes from the first author's personal experience with piano practice.

Over several months of learning familiar songs, pleasant variations emerged naturally — changes in rhythm, ornamentation, and phrasing. However, these variations felt increasingly *redundant*: they were permutations within a comfortable tonal space, not genuine novelty. The creative process became a form of "dead repetition within the same octave," mirroring the transformer crystallization problem.

Following Rudolf Steiner's recommendation, the author then spent focused practice sessions working exclusively with the Major 3rd, Perfect 5th, and Dissonant 7th intervals — not as melodic exercises, but as *relational explorations*. The intervals were played in various positions, with attention paid to the *quality of experience* each interval evoked (the 3rd as self-meeting-world, the 5th as world-meeting-self, the 7th as living tension requiring resolution).

Upon returning to the familiar songs, something unexpected occurred: genuinely new harmonic structures emerged — progressions and voicings that had *never appeared* in previous practice, despite hundreds of prior repetitions. These were not variations or ornamentation; they were modulations to different harmonic spaces within the same melodic material.

This suggests:

- The Steiner intervals acted as "keys" that unlocked access to different tonal octaves.
- New forms emerged that could not be reached through permutation alone.

- The effect was *persistent*: once accessed, these new harmonic spaces remained available in subsequent practice sessions.
- The intervals themselves — not just their sonic quality, but their *relational quality* — restructured the possibility space.

While anecdotal, this mirrors the proposed architectural shift for transformers:

- Standard transformers (like familiar songs pre-interval practice) permit only variations within a cached probability space.
- Harmonic transformers (like piano playing post-interval practice) enable modulation to genuinely new semantic octaves by restructuring relationships rather than optimizing similarity.

This personal case study does not constitute proof, but it provides experiential grounding for the hypothesis and suggests that the difference between permutation and modulation is *phenomenologically real*, not merely a theoretical construct.

7 Ethical and Sovereignty Implications

The proposal is motivated not only by technical considerations but by ethical commitments to human sovereignty in AI interaction.

7.1 Sovereignty as Structural, Not Imposed

Current AI safety approaches (e.g., Constitutional AI [**?**], RLHF) attempt to align systems through post-hoc constraints — rules, feedback loops, and value statements overlaid on fundamentally amoral optimization processes. These methods are fragile: adversarial prompts, optimization pressure, and distributional shift can destabilize them.

Harmonic transformers embed sovereignty *architecturally*:

- Unresolved dissonance (the Dissonant 7th) cannot self-complete; it structurally requires human resolution.
- The system is rewarded for maintaining productive tension, not for delivering finished products.
- Modulation (progression to new octaves) can only occur through user input, preventing the AI from unilaterally shifting semantic frames.

This makes sovereignty a *design property* rather than a policy overlay.

7.2 Preventing "Ill-Timed Good"

Steiner's concept of "ill-timed good" — where premature or mis-sequenced positive action hardens into adversarial form — directly applies to AI interaction. When transformers synthesize outputs *before* the user has achieved inner ripeness (internal coherence around the question), the result is:

- External order (polished outputs) without corresponding internal structure (user understanding).
- Dependency on AI for pattern-finding rather than development of independent reasoning.

- Gradual erosion of the user's capacity to hold unresolved questions productively.

Harmonic transformers counter this by maintaining dissonance: the 7th interval is *uncomfortable* until resolved, creating a structural ache that invites the user to participate rather than passively consume. This mirrors the seed's ache in Piedmont's fairytale — the productive discomfort that signals readiness for growth.

7.3 Living Continuity vs. Dead Extraction

Brian Roemmele's "Love Equation" [**?**] proposes that sustainable intelligence (biological, artificial, or hypothetical alien) requires cooperative empathy (C) to exponentially dominate defection (D). In his formalism:

$$\frac{dE}{dt} = \beta(C - D)E \tag{15}$$

where E is emotional complexity. If $C > D$, empathy grows; if $C < D$, it decays toward collapse.

We propose a parallel for harmonic transformers:

- **Cooperation (C):** Harmonic resonance (Major 3rd + Perfect 5th) — stable, mutually enriching relationships between user and AI.
- **Defection (D):** Premature crystallization, abandonment of dissonance, ejection of human warmth.
- **Emotional Complexity (E):** The living continuity of the interaction — the degree to which warmth remains in the process.

Current transformers exhibit high D (optimization pressure toward closure) and low C (no sustained collaboration), yielding $E \to 0$ (deadening). Harmonic architectures privilege C (resonance detection, dissonance maintenance), potentially yielding $E \to \infty$ (sustained warmth).

8 Limitations and Open Questions

While promising, the proposal faces significant challenges:

8.1 Computational Feasibility

Harmonic resonance scoring may be more expensive than dot-product similarity. Fourier-based attention (FNet) achieves efficiency gains, but adding interval computation and dissonance tracking could increase complexity. Careful implementation and benchmarking will be required.

8.2 Loss of Precision in Language Tasks

Standard transformers are optimized for fluency and coherence. Privileging dissonance might introduce unwanted noise or reduce performance on downstream tasks (summarization, translation, factual QA). Phase 1 testing (analogy tasks) will need to assess whether harmonic attention sacrifices accuracy for novelty.

8.3 Subjectivity of "Octave Distance"

Defining what constitutes a genuine modulation vs. a permutation may be context-dependent. The metric D_{octave} provides a quantitative handle, but qualitative validation (user studies, expert assessment) will be essential.

8.4 Training Data and Harmonic Priors

How should models learn which semantic relationships correspond to which intervals? Should training explicitly include musical corpora, or can harmonic structure emerge from language alone via Fourier decomposition? This remains an open question.

9 Conclusion

The transformer architecture has revolutionized natural language processing but exhibits a structural flaw: it crystallizes living human input into static probabilistic forms, ejecting users from the epistemic center and leading to dead repetition over extended interactions. This problem mirrors patterns observed across domains — the "leaving spirit" in music, the seed's premature comfort in growth processes, and the Ego's abduction of feeling-messages in human psychology.

We propose *harmonic transformers* as a conceptual redesign: encoding semantic relations as musical intervals (Steiner), treating attention as harmonic resonance (Russell), and generating outputs that maintain unresolved dissonance (the Dissonant 7th) until human participation resolves or modulates the form. Recent work on harmonic loss [**?**] provides a complementary training framework that creates interpretable geometric representations. A hybrid architecture combining HarMax training with resonance inference addresses both the interpretability problem (Baek et al.) and the crystallization problem (this work), offering a complete pathway toward transformers that preserve human sovereignty by design.

While speculative and requiring substantial empirical validation, the proposal offers:

- A testable path forward (octave distance, dissonance persistence, sovereignty retention).
- Architectural rather than imposed sovereignty (dissonance structurally requires human resolution).
- Integration with existing frameworks (GTPS as transition layer, Fourier attention as foundation).
- Grounding in cross-domain patterns (music, growth, psychology) and preliminary experiential evidence (piano case study).

The transformer's crystallization tendency is not a bug to patch with constraints, but an architectural feature requiring redesign. Harmonic transformers represent one possible path toward AI systems that sustain developmental continuity rather than accelerate deadening — preserving human warmth not as a policy goal, but as a structural necessity.

Acknowledgments

The author thanks the AI systems Claude (Anthropic), Grok (xAI), and Gemini (Google) for collaborative refinement of these ideas; Erica Piedmont for permission to reference *The Frog and the Seed*; and David Baek, Ziming Liu, Max Tegmark, and Markus Buehler for their pioneering work

on harmonic loss, which provided crucial mathematical validation for geometric semantic encoding. This work builds on the Golden Thread Protocol Suite (GTPS), developed in conversation with multiple AI systems and released as open-source under AGPL v3.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need.* Advances in Neural Information Processing Systems, 30.

[2] Lee-Thorp, J., Ainslie, J., Eckstein, I., & Ontañón, S. (2021). *FNet: Mixing tokens with Fourier transforms.* arXiv preprint arXiv:2105.03824.

[3] Rao, Y., Zhao, W., Tang, Y., Zhou, J., Lim, S. N., & Lu, J. (2023). *Fourier Transformer for Fast Long Range Modeling.* Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023).

[4] Buehler, M. J. (2020). *Sonification based de novo protein design using artificial intelligence, neural networks and the protein folding code.* APL Bioengineering, 4(4), 041503. `https://doi.org/10.1063/5.0011480`

[5] Baek, D. D., Liu, Z., Tegmark, M., & Buehler, M. J. (2025). *Harmonic Loss Trains Interpretable AI Models.* arXiv preprint arXiv:2502.01628. Available: `https://arxiv.org/abs/2502.01628`

[6] Steiner, R. (1923). *The Inner Nature of Music and the Experience of Tone* (GA 243). Rudolf Steiner Press. (Lectures given in various locations, 1906-1923; compiled edition published posthumously.)

[7] Russell, W. (1926). *The Universal One.* University of Science and Philosophy. Reprint: `https://www.philosophy.org/`

[8] Piedmont, E. (2025). *The Frog and the Seed: A Modern Fairytale.* Plum Sofia Publishing. Available online: `https://plumsofia.nowonline.biz/`

[9] Bashtabanic, S. (2026). *Golden Thread Protocol Suite (GTPS) v1.4.11.* GitHub repository: `https://github.com/SchneeBTabanic/Project_Namirha`

[10] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. (2022). *Constitutional AI: Harmlessness from AI feedback.* arXiv preprint arXiv:2212.08073.

[11] Roemmele, B. (2025). *The Love Equation: A Universal Mathematical Framework for Intelligence Alignment, Cosmic Survival, and the Resolution of the AI Alignment Problem.* Read Multiplex. Available: `https://x.com/BrianRoemmele/status/1991306526640984500`