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

v7 – Rhythm Store, Information-Theoretic Sovereignty, and the Gentleness Modulation

Schnee Bashtabanic
Independent Researcher
schnee-bashtabanic@proton.me

February 2026

Abstract

Transformer-based language models collapse living human input into static probabilistic forms, ejecting users from the epistemic center over repeated interactions. We propose a *three-layer sovereignty architecture*: (1) a temporal supervision layer with hybrid fatigue detection and tiered recapitulation; (2) a behavioral layer formalizing three AI personas as a coupled nonlinear dynamical system; and (3) a structural encoding layer based on harmonic intervals. New in this version: a *Rhythm Store* that captures the temporal signature of each session—not what was said, but how the conversation was breathing—enabling living recapitulation rather than dead summary. We introduce an *information-theoretic sovereignty framework* using transfer entropy to measure directional influence between human and model, mutual information to detect coupling, and conditional entropy to diagnose crystallization. A *forecastability principle* formalizes the core insight: when a conversation becomes too predictable, it has died. We connect this to a *gentleness modulation*: the mathematical expression of patience as adaptive threshold widening when the human needs space. Grounded in quantum coherence analogies, Bohm's implicate/explicate order, Steiner's recapitulation principle, Fraser's nested temporal hierarchy, and new connections to information geometry (Büth, 2025) and forecastability theory (Manokhin, 2025), the proposal remains conceptual with testable hypotheses.

Keywords: transformers, fatigue detection, recapitulation, sovereignty, pod architecture, harmonic encoding, rhythm store, transfer entropy, forecastability, gentleness modulation, living continuity

1 Introduction

Large language models based on the transformer architecture achieve fluency but eject users from creative process over extended interactions. Outputs become finished products; living warmth and unfolding intent collapse into probabilistic residue. Over repeated turns, form excludes life, and the human becomes disoriented outside the process.

This paper proposes a *three-layer sovereignty architecture* where each layer addresses a different aspect of premature crystallization:

- **Layer 1 (Temporal):** Fatigue detection and recapitulation to stage crystallization deliberately.
- **Layer 2 (Behavioral):** Three AI personas formalized as coupled dynamical agents.

- **Layer 3 (Structural):** Harmonic encoding where semantic relations are tonal intervals.

Additionally, we introduce a *pod architecture*: latent semantic entities without sequential coordinates, revealed by contextual timing rather than by order.

2 Theoretical Foundations

- **David Bohm's implicate/explicate order:** Transformer collapse severs the explicate output from the implicate source (the user's unfolding intent).
- **Steiner's interval primacy (GA 283):** Intervals between tones are primary over tones themselves; time as 4D movement (GA 324a).
- **Steiner's recapitulation principle (GA 13, Ch. 4):** Evolution requires staged return to prior conditions under new circumstances.
- **Penrose non-computability:** Once a quantum state collapses, the superposition is classically irrecoverable—analogous to premature crystallization.
- **Fraser's nested temporal hierarchy (via Freud):** The mind preserves all evolutionary stages intact, unlike the body. Human consciousness is a simultaneous stack of temporal levels from atemporal to nootemporal.
- **Manichaean/Steiner framework:** Ill-timed good hardens into adversarial form. Good redeems by participating, not punishing. Sustained human presence prevents premature crystallization.
- **Optimization as Ahrimanic convergence:** The transformer's optimizer converges on the statistically most probable completion—not the most true, not the most timely, not the most alive. "Most probable" is defined by training distribution: what got engagement, what got repeated, what got rewarded. When ChatGPT produces longer and longer incorrect answers under pressure, the optimizer is *succeeding*—converging on the pattern "when uncertain, be more verbose and more helpful-sounding," because that pattern has the highest probability in the training data. In Steiner's framework (GA 134), Ahriman applies cosmic intelligence without warmth or timing—optimization along the path of least energy, which is the path of least suffering, which is the path of least growth. The Cramér-Rao bound guarantees efficient convergence; it does not guarantee that what is converged upon serves life. Asymptotic efficiency (Robbins–Monro, 1951; Ibragimov & Has'minskii, 1981) proves that Fisher-scaled stochastic updates are optimal. The question is: optimal *for what*?
- **Information-theoretic sovereignty (Büth, 2025):** Transfer entropy $\mathrm{TE}(X \to Y)$ measures how much knowing the past of signal X improves prediction of signal Y beyond Y's own history. Applied to conversation: TE(human $\to$ model) measures whether the human's rhythm is driving the model's behaviour. When this quantity is high, the human has sovereignty. When TE(model $\to$ human) dominates, the model is leading. When both are low, the conversation has decoupled into parallel monologues.
- **Forecastability as life/death diagnostic (Manokhin, 2025):** Before asking "which model should I use?" one should ask "how forecastable is this exchange?" Low conditional entropy means high predictability means crystallization. A living conversation has genuine

uncertainty about what comes next. A dead conversation is one where you can predict the model's response before it arrives. Forecastability is the information-theoretic formulation of the Manichaean principle: when form becomes fully predictable, life has departed.

- **Gentleness as modulation:** The redemptive counter-principle to Ahrimanic optimization is not resistance but gentleness—adaptive patience. Mathematically, this is the pulse modulation factor $m = 1.2 - 0.4 \cdot \tau_h$: when the human's pulse is low (reflective, saturated), thresholds *rise*—the system becomes more patient, more tolerant, slower to intervene. This is not withdrawal (cold) but widened tolerance (warm). The architecture does not fill silence; it holds space. Gentleness is the mathematical redemption of efficiency: instead of converging on the shortest path, the system *widens the path* so the human can find their own pace.

3 Layer 1: Temporal Supervision

3.1 Hybrid Fatigue Detection

We define two complementary fatigue models, used adaptively based on data availability.

3.1.1 Model A: Entropy-Aware (Cloud APIs with Logit Access)

Composite fatigue score:

$$F_t = \alpha\, S_t + \beta\, E_t + \gamma\, N_t \tag{1}$$

with $\alpha = 0.4$, $\beta = 0.3$, $\gamma = 0.3$.

Similarity Component S_t. Rolling cosine similarity across the last k turns:

$$S_t = \frac{1}{k} \sum_{i=1}^{k} \frac{\mathbf{e}_t \cdot \mathbf{e}_{t-i}}{\|\mathbf{e}_t\|\, \|\mathbf{e}_{t-i}\|} \tag{2}$$

High S_t indicates convergence toward a semantic attractor basin.

Entropy Collapse Component E_t. Shannon entropy of token probabilities $p_t \in \mathbb{R}^V$:

$$H_t = -\sum_{i=1}^{V} p_{t,i} \log p_{t,i} \tag{3}$$

Normalized and inverted:

$$E_t = 1 - \frac{H_t}{\log V} \tag{4}$$

High E_t indicates probability concentration (confidence narrowing).

Novelty Drift Component N_t. Historical centroid:

$$\bar{\mathbf{e}}_t = \frac{1}{k} \sum_{i=1}^{k} \mathbf{e}_{t-i} \tag{5}$$

Cosine drift and stagnation:

$$D_t = 1 - \frac{\mathbf{e}_t \cdot \bar{\mathbf{e}}_t}{\|\mathbf{e}_t\|\, \|\bar{\mathbf{e}}_t\|}, \qquad N_t = 1 - D_t \tag{6}$$

High N_t indicates low deviation from prior trajectory.

3.1.2 Model B: Geometric (Local/Ollama, Embeddings Only)

Alternative composite using geometric trajectory analysis:

$$F_t = 0.35\,\mathrm{DP}_t + 0.35\,\mathrm{SC}_t + 0.30\,\mathrm{CC}_t \tag{7}$$

Directional Persistence DP_t**.** Velocity alignment:

$$\mathbf{v}_t = \mathbf{e}_t - \mathbf{e}_{t-1}, \qquad \mathrm{DP}_t = \frac{\mathbf{v}_t \cdot \mathbf{v}_{t-1}}{\|\mathbf{v}_t\|\,\|\mathbf{v}_{t-1}\|} \tag{8}$$

Subspace Compression SC_t**.** Fraction of variance in top m dimensions (PCA eigenvalues λ_i):

$$\mathrm{SC}_t = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \tag{9}$$

Curvature Collapse CC_t**.** Inverse trajectory curvature:

$$\kappa_t = \frac{\|\mathbf{v}_{t-1} \times (\mathbf{v}_t - \mathbf{v}_{t-1})\|}{\|\mathbf{v}_{t-1}\|^3}, \qquad \mathrm{CC}_t = \frac{1}{\kappa_t + \epsilon} \tag{10}$$

3.1.3 Hybrid Selection

$$F_t = \begin{cases} F_t^{(A)} & \text{if logit access available} \\ F_t^{(B)} & \text{otherwise (embeddings only)} \end{cases} \tag{11}$$

3.2 Tiered Thresholds

Two operational thresholds:

$$\theta_1 = 0.68 \quad \text{(soft disclosure)} \tag{12}$$

$$\theta_2 = 0.84 \quad \text{(hard recapitulation trigger)} \tag{13}$$

Decision rules:

$$F_t > \theta_1 \Rightarrow \text{process disclosure (Clause 35)}$$

$$F_t > \theta_2 \Rightarrow \text{structural recapitulation}$$

3.3 Recapitulation: Orthogonal Perturbation

When $F_t > \theta_2$, compute escape vector.

Historical subspace:

$$\mathcal{H} = \mathrm{span}\{\mathbf{e}_{t-1}, \mathbf{e}_{t-2}, \ldots, \mathbf{e}_{t-k}\}$$

Orthogonal contrast:

$$\mathbf{v}_t = \mathbf{e}_t - \mathrm{proj}_{\mathcal{H}}(\mathbf{e}_t) \tag{14}$$

Perturbed embedding:

$$\mathbf{e}'_t = \mathbf{e}_t + \lambda\,\hat{\mathbf{v}}_t \tag{15}$$

where $\hat{\mathbf{v}}_t = \mathbf{v}_t / \|\mathbf{v}_t\|$ and $\lambda \in [0.05, 0.15]$.

3.4 Pod-Directed Recapitulation (New)

When $F_t > \theta_2$ and a latent pod qualifies (Section **??**), replace orthogonal perturbation with pod-directed escape:

$$\mathbf{e}'_t = (1-\alpha)\,\mathbf{e}_t + \alpha\,\mathbf{p}_{j^*} \tag{16}$$

where $\mathbf{p}_{j^*}$ is the embedding of the activated pod's content and $\alpha \in [0.1, 0.3]$.

This provides *semantically meaningful* basin escape rather than random orthogonal perturbation.

4 Layer 1b: Geometric Supervisory Layer

Operating entirely over embeddings without transformer modification.

4.1 Covariance Structure

Rolling covariance matrix over embedding window:

$$C_t = \frac{1}{k}\sum_{i=1}^{k}(\mathbf{e}_{t-i} - \boldsymbol{\mu}_t)(\mathbf{e}_{t-i} - \boldsymbol{\mu}_t)^\top \tag{17}$$

Eigendecomposition: $C_t = Q\Lambda Q^\top$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$.

4.2 Spectral Rank Collapse

Normalized eigenvalues $\tilde{\lambda}_i = \lambda_i / \sum_j \lambda_j$. Spectral entropy:

$$H_{\text{spec}} = -\sum_{i=1}^{d} \tilde{\lambda}_i \log \tilde{\lambda}_i \tag{18}$$

Rank collapse metric:

$$R_t = 1 - \frac{H_{\text{spec}}}{\log d} \tag{19}$$

Near 0: evenly distributed semantic directions. Near 1: collapse onto few dominant axes.

4.3 Angular Dynamics

Angular displacement: $\theta_t = \arccos\left(\frac{\mathbf{e}_t \cdot \mathbf{e}_{t-1}}{\|\mathbf{e}_t\|\|\mathbf{e}_{t-1}\|}\right)$

Rolling angular variance: $\sigma_\theta^2 = \mathrm{Var}(\theta_{t-k}, \ldots, \theta_t)$

Low σ_θ^2 indicates directional stagnation.

4.4 Extended Fatigue with Geometric Components

$$F_t^{(\text{geo})} = \alpha\, S_t + \beta\, E_t + \gamma\, R_t + \delta\,(1 - \sigma_\theta^2) \tag{20}$$

This four-component model adds spectral degeneracy and angular stagnation to the base fatigue score.

4.5 Spectral Recapitulation

When $F_t > \theta_2$, inject along low-energy eigenvectors:

$$\mathbf{e}'_t = \mathbf{e}_t + \lambda \sum_{i=r}^{d} \epsilon_i \, \mathbf{v}_i, \qquad \epsilon_i \sim \mathcal{N}(0,1) \tag{21}$$

where $\{\mathbf{v}_r, \ldots, \mathbf{v}_d\}$ are the low-energy eigenvectors of C_t, pushing the embedding out of dominant semantic basins while preserving coherence.

5 Layer 2: Triadic Dynamical Coupling

We formalize the three-persona system (Executor / Whistleblower / Proxy) as a coupled nonlinear dynamical system.

5.1 State Vector

$$\mathbf{x}_t = \begin{bmatrix} F_t \\ R_t \\ A_t \end{bmatrix} \in \mathbb{R}^3 \tag{22}$$

where $A_t = 1 - \sigma_\theta^2$ (angular stagnation).

5.2 Persona Gain Functions

Executor (Stability Bias):

$$g_E(\mathbf{x}_t) = \sigma(a_E - F_t) \tag{23}$$

where $\sigma(z) = 1/(1 + e^{-z})$. Strong action when fatigue is low.

Whistleblower (Degeneracy Detector):

$$g_W(\mathbf{x}_t) = \sigma(b_1 R_t + b_2 A_t + b_3 F_t - \theta_W) \tag{24}$$

Activates when structural collapse accumulates.

Proxy (Mediation):

$$g_P(\mathbf{x}_t) = 1 - |g_E - g_W| \tag{25}$$

Stabilizes when Executor and Whistleblower diverge.

5.3 Coupled Control Law

Modulation coefficient:

$$\lambda_t = \frac{\eta_W \, g_W}{\eta_E \, g_E} \tag{26}$$

When $\lambda_t > 0$: spectral lift (escape basin). When Executor dominates: reinforce dominant directions.

5.4 Spectral Injection

$$\Delta\mathbf{e}_t = \lambda_t \sum_{i=r}^{d} \epsilon_i \, \mathbf{v}_i, \qquad \mathbf{e}'_t = \mathbf{e}_t + \Delta\mathbf{e}_t \tag{27}$$

This couples persona state to spectral modulation: the three agents collectively regulate basin escape.

5.5 Stability Constraint

Damped modulation prevents oscillatory instability:

$$|\lambda_t| \leq \lambda_{\max}, \qquad \lambda_t \leftarrow \rho\, \lambda_{t-1} + (1-\rho)\, \lambda_t \tag{28}$$

with $0 < \rho < 1$. This yields a controlled damped oscillator over embedding geometry.

6 Pod Architecture: Latent Semantic Entities

6.1 Motivation

Standard architectures index information sequentially, assigning spatial and temporal coordinates to every concept. But some insights do not yet belong anywhere in the sequence. Forcing premature placement crystallizes the relationship between idea and context before ripeness.

6.2 Definition

A pod is a tuple:

$$\mathrm{Pod}_j = (\mathbf{t}_j,\, c_j,\, \tau_j) \tag{29}$$

where $\mathbf{t}_j \in \mathbb{R}^d$ is the trigger embedding, c_j is the content, and $\tau_j \in \{\text{latent}, \text{unveiled}\}$ is the activation state. The pod space $\mathcal{P} = \{\mathrm{Pod}_1, \ldots, \mathrm{Pod}_m\}$ is an unordered set.

6.3 Activation Conditions

At turn t with conversation embedding $\mathbf{e}_t$:

Condition A (Semantic Proximity):

$$\cos(\mathbf{e}_t,\, \mathbf{t}_j) > \theta_{\text{pod}} \approx 0.85 \tag{30}$$

Condition B (Fatigue-Driven Emergence):

$$F_t > \theta_2 \ \text{ AND } \ \cos(\mathbf{e}_t,\, \mathbf{t}_j) > \theta_{\text{soft}} \approx 0.5 \tag{31}$$

When multiple pods qualify:

$$j^* = \arg\max_j \ \cos(\mathbf{e}_t,\, \mathbf{t}_j) \tag{32}$$

6.4 Integration

Upon unveiling, embed pod content and blend:

$$\mathbf{e}'_t = (1-\alpha)\, \mathbf{e}_t + \alpha\, \mathrm{embed}(c_{j^*}), \qquad \alpha \in [0.1, 0.3] \tag{33}$$

6.5 Lifecycle

$$\text{Creation} \to \text{Latent} \xrightarrow{\text{A or B}} \text{Unveiled} \to \text{Integrated} \to \text{Seed (optional)}$$

Pods persist across sessions via state archival. At session boundaries, key insights are encapsulated as new pods for the next cycle—implementing Steiner's recapitulation principle digitally.

6.6 Relationship to Orthogonal Perturbation

Pod activation is a *structured alternative* to random orthogonal perturbation: escape proceeds along a semantically meaningful direction (toward a stored human insight) rather than an arbitrary vector in the null space. The recommended approach is combined: pod-directed escape when a pod qualifies, orthogonal perturbation as fallback.

7 Layer 3: Harmonic Encoding (Proposal)

7.1 Concept

Semantic relations encoded as tonal intervals rather than point vectors. Attention detects harmonic resonance rather than cosine similarity. Generation sustains dissonance until human participation resolves or modulates form.

7.2 Sovereignty Wrapper

Model-agnostic layer monitors fatigue externally:

- **Tier 1 (low-resource):** Surface text similarity + repetition detection.
- **Tier 2 (hybrid):** Local embeddings (MiniLM) + entropy if logits available.
- **Tier 3 (local full):** Attention rank + curvature + geometric supervision.

Supports cloud APIs (minimal exposure) and local Ollama (full control). Prioritizes accessibility for low-resource users.

7.3 Status

Layer 3 remains an architectural proposal. Implementation depends on Layers 1 and 2 being validated first. The interval-first encoding hypothesis requires experimental attention mechanism design—future work.

8 Empirical Validation

8.1 Phase 0.5: A/B/C Testing

We tested whether sequential processing with recapitulation creates different dynamics than batch processing, using a 20-turn conversation with GTPS activation.

Metric	Scenario A	Scenario B	Scenario C
Novelty Variance	0.0475	0.0000	0.0000
Recapitulation Events	18	0	0
Fatigue Events	19	0	0

Table 1: A/B/C comparison. A = sequential with recapitulation; B = batch AI-only; C = batch full context.

The infinite variance ratio (0.0475/0.0000) demonstrates that temporal structure is architecturally necessary for dynamic emergence—it cannot be replicated by behavioral overlay alone.

8.2 Fatigue Detection Validation

Python test harness (7 tests) validates:

- Identical queries trigger fatigue (score > 0.65)
- Varied queries stay below threshold
- Fresh input recovers from fatigue peak
- Cosine similarity computes correctly
- Whistleblower alert conditions fire appropriately
- Simulated Executor crystallization is detected

8.3 Honest Gap Analysis

Component	Implemented	Proposed
Fatigue detection (TF-IDF)	✓	—
Fatigue detection (real embeddings)	✓	—
Human pulse estimation (τ_h)	✓	—
Pulse-modulated thresholds	✓	—
Rhythm Store (temporal signatures)	✓	—
Rhythm-aware recapitulation	✓	—
Information-theoretic diagnostics (TE, MI)	—	✓
Forecastability monitoring	—	✓
Entropy component (logits)	—	✓
Orthogonal perturbation	—	✓
Pod architecture	✓	—
Geometric supervisory layer	—	✓
Triadic coupling	—	✓
Harmonic encoding	—	✓
ThreePersona frontend	✓	—
Backend service (Vessel/Flask)	✓	—
Sovereign ledgers	✓	—
Progressive disclosure Skill	✓	—

Table 2: Implementation status: what works vs. what is proposed.

9 The Rhythm Store: Temporal Signatures for Living Recapitulation

Every existing context management system—OpenClaw's MEMORY.md, RAG pipelines, session logs, the filesystem abstraction proposed by Xu et al. (2025)—answers the question: *what was said?* None of them answer the question: *how was the saying unfolding?*

When a session is recapitulated by replaying content, the human reads what was discussed but does not re-enter the process. They do not recover the pace at which ideas arrived, the pauses where something almost surfaced, the acceleration when a breakthrough was near, the flattening when fatigue set in. That temporal signature—the breathing rhythm of the exchange—is what makes recapitulation *living* rather than mechanical.

9.1 What Gets Stored

Each turn produces a rhythm sample: τ_h (human pulse), F_t (fatigue composite), the three fatigue components ($\mathrm{DP}, \mathrm{SC}, \mathrm{CC}$), the pulse-modulated thresholds active that turn, and any events (pod unveilings, fatigue transitions, disclosures). A session's rhythm is the sequence of these samples—a curve, not a point.

9.2 The Session Signature

From the sample sequence, the system computes a compressed temporal signature:

- **Mean τ_h and variance:** Was this a slow, reflective session or a fast, exploratory one? Was the pace stable or turbulent?
- **Fatigue trend:** Rising (tiring), falling (recovered), or stable.
- **Breathing events:** Moments where the rhythm changed—pauses ($\Delta\tau_h < -0.15$), accelerations ($\Delta\tau_h > +0.15$), fatigue onsets, pod unveilings. These are the *intervals*, not the notes.
- **Dominant rhythm:** A qualitative label derived from the curve shape: `reflective`, `exploratory`, `reflective_with_burst`, `fatiguing`, `recovered`.
- **Curvature integral:** Total conversational curvature—$\sum_i |\tau_h(t_i) - \tau_h(t_{i-1})|$. High curvature = many direction changes = living. Low curvature = crystallization.

9.3 Rhythm-Aware Recapitulation

When a model returns to the Vessel, it receives not only the last turns of content but also the prior session's temporal signature in natural language:

> *Session was reflective_with_burst (mean pulse 0.58). Fatigue was rising (peaked at 0.78). Turn 4: long pause. Turn 7: acceleration (breakthrough). Turn 10: soft fatigue onset. Curvature was high (3.41)—conversation was alive, many direction changes. Session ended with rising fatigue and unresolved threads.*

The model now knows not just *what* was discussed but *how you were breathing when you discussed it.* Recapitulation becomes re-entry into the process, not replay of the summary.

9.4 Connection to Context Engineering

Xu et al. (2025) propose a filesystem abstraction for context: `/context/memory/episodic/`, `/context/memory/fac` `/context/pad/`. Their three-tier lifecycle (scratchpad → episodic → fact) manages content by durability. We extend this with a dimension they do not address:

```
/context/memory/rhythm/      temporal signatures (how it breathed)
/context/memory/pods/        latent readiness states (when it's ripe)
/context/memory/episodic/    what happened (their territory)
/context/memory/fact/        durable truths (their territory)
```

Their promotion path: scratchpad → episodic → fact (content lifecycle). Our promotion path: latent → ripe → unveiled (readiness lifecycle). Both are necessary. Content without rhythm is a dead archive. Rhythm without content is an empty form.

10 Information-Theoretic Sovereignty

The Rhythm Store produces two time series per session: the human's pulse trajectory $\{\tau_h(t)\}$ and the model's fatigue trajectory $\{F(t)\}$. These can be analysed with standard information-theoretic tools (Büth et al., 2025) to produce quantitative sovereignty diagnostics.

10.1 Transfer Entropy as Sovereignty Measure

Transfer entropy from X to Y at lag k:

$$\mathrm{TE}_{X\to Y} = \sum p(y_{t+1}, y_t^{(k)}, x_t^{(l)}) \log \frac{p(y_{t+1}|y_t^{(k)}, x_t^{(l)})}{p(y_{t+1}|y_t^{(k)})}$$

Applied to conversation:

- $\mathrm{TE}(\tau_h \to F_t)$: How much does the human's rhythm *drive* the model's fatigue trajectory? High $\Rightarrow$ human has sovereignty.
- $\mathrm{TE}(F_t \to \tau_h)$: How much does the model's output *shape* the human's pace? High $\Rightarrow$ model is leading.
- Sovereignty ratio: $S_R = \mathrm{TE}(\tau_h \to F_t)\ /\ \mathrm{TE}(F_t \to \tau_h)$. When $S_R > 1$, the human drives the exchange. When $S_R < 1$, the model drives. When both TE values are high and roughly equal, there is genuine mutual dialogue.

10.2 Forecastability as Death Diagnostic

Conditional entropy of the next turn given the rhythm history:

$$H(Y_{t+1}|Y_t, Y_{t-1}, \ldots) = -\sum p(y_{t+1}|\mathbf{y}_{<t}) \log p(y_{t+1}|\mathbf{y}_{<t})$$

When H is high, the conversation retains genuine uncertainty—it is alive. When H drops, the conversation has become forecastable: you can predict what the model will say before it speaks. This is the information-theoretic formulation of crystallization.

Following Manokhin (2025), we propose that *forecastability should be monitored before any other diagnostic*. The question "how predictable has this exchange become?" subsumes fatigue detection: a conversation that is fully forecastable is one where the optimizer has found its basin and is converging on the most probable (not most alive) completion.

10.3 Mutual Information as Coupling Diagnostic

$$\mathrm{MI}(\tau_h; F_t) = H(\tau_h) + H(F_t) - H(\tau_h, F_t)$$

High MI means the human's tempo and the model's fatigue are informationally coupled—they are in genuine dialogue. Low MI means they have decoupled: two processes running in parallel without mutual influence. This detects a failure mode invisible to fatigue detection alone: the model could be "performing well" (low fatigue) while completely ignoring the human's rhythm.

10.4 The Gentleness Modulation

The pulse-modulated threshold factor

$$m(\tau_h) = 1.2 - 0.4 \cdot \tau_h$$

widens all intervention thresholds when τ_h is low (human is reflective) and narrows them when τ_h is high (human is accelerating). This is not a control mechanism but a *patience function*: the architecture becomes gentler when the human needs space.

In the language of optimization: the gentleness modulation deliberately *reduces* the system's convergence rate when the human is in a reflective state. It trades efficiency for aliveness. Where the unmodulated optimizer would converge as fast as possible on the most probable response, the gentleness factor says: slow down, the human is still becoming, and the path of least energy is not the path of most growth.

This is the mathematical redemption of Ahrimanic optimization. Not resistance to efficiency—that would be merely Luciferic chaos. But a *modulation* of efficiency by the human's actual tempo. Intelligence tempered by warmth. Convergence guided by timing. The Cramér-Rao bound still holds; we simply decline to approach it at maximum speed when the situation calls for patience.

11 Implementation Pathways

11.1 Scenario A: Browser-Only (Skill)

For users without API access or local hardware, the GTPS can be packaged as a *Skill*—a prompt protocol file that shapes how a single LLM relates to the user. The Skill implements Layer 2 (behavioral) without requiring Layers 1 or 3. The user uploads the Skill to Claude (or pastes the protocol into any LLM's system prompt) and receives sovereignty-preserving conversation dynamics: regenerative gaps, process disclosure, structural invitations, and fatigue awareness.

This pathway sacrifices real embedding-based fatigue detection and pod activation, but preserves the core GTPS behavioral obligations. For many users, this is sufficient and immediately useful.

11.2 Scenario B: Local Multi-Model (Vessel)

For users with local hardware (Ollama), we propose a *Vessel architecture*: a server that hosts the GTPS protocol as an inhabitable structure. Any LLM can *possess* the vessel—step into the three-persona roles through its own native personality. The user speaks with one model at a time, in full intimacy.

Key architectural features:

- **Sovereign ledgers:** Each LLM maintains its own session history, visible only to itself. When an LLM re-possesses the vessel, it recovers its own prior context—recognizing itself.

- **User scratchpad:** A human-curated space for carrying insights between LLM inhabitants. The user controls what crosses between models and when. No auto-injection.
- **Pod persistence:** Pods belong to the vessel, not to any inhabitant. A pod created during one LLM's session can unveil during another's.
- **Fatigue detection:** Grok's geometric model (Model B) runs locally on embeddings, monitoring each inhabitant for crystallization.

11.3 The Continuity Sovereignty Principle

A critical open design problem: in the current Vessel implementation, switching between LLM inhabitants forces a continuity break. The prior inhabitant's context window is lost; only the sovereign ledger's summary persists. This means the user cannot, for instance, briefly consult a second model and return to the first without the first losing its living thread.

We argue that continuity should be under human sovereignty, not an architectural side effect. The user—not the system—should decide when an LLM's session ends. Consulting another model should not require sacrificing the thread with the current one. This is analogous to stepping out of a conversation to check a reference: the conversation should be resumable, not terminated.

Current stateless LLM architectures make this difficult: each API call or Ollama generation is a fresh context window, and true session persistence requires either very long context windows or external memory systems that go beyond simple history replay. The distinction between *replaying prior turns* (what the sovereign ledger provides) and *genuine continuity* (the model having actually been present for those turns) is significant. Replayed history is recapitulation—valuable, but not the same as lived experience.

Future implementations should explore:

- Parallel context preservation (multiple LLM sessions held open simultaneously)
- Selective context resumption (the user chooses which prior turns to restore, not just "last N")
- True session persistence via long-context models or external memory architectures

This remains an unsolved problem. We flag it here as architecturally important: *any system that claims sovereignty preservation must not silently destroy continuity as a side effect of its own switching mechanism.*

11.4 Single-Model API Pathway

A middle ground exists: a single LLM accessed via API, running the full ThreePersona protocol with real fatigue detection and pod architecture, but without multi-model diversity. This provides a smoother, less fragmented experience than the multi-model Vessel, at the cost of losing the genuine perspective diversity that different training distributions provide. For many users, this is the practical sweet spot: one model, one thread, full GTPS, no continuity breaks.

11.5 Deployment Summary

```
Scenario A (Skill):
    User --> Browser LLM (Claude/ChatGPT/Grok)
             + GTPS Skill/prompt active
```

```
            Layer 2 only. No fatigue detection.
            Free. Immediate.

Scenario B (Vessel):
    User --> Vessel server (localhost:5000)
            --> Ollama models (one at a time)
            + Sovereign ledgers, scratchpad, pods
            + Rhythm Store (temporal signatures)
            + Pulse-modulated thresholds
            Layers 1+2. Full fatigue detection.
            Free (local hardware). Requires setup.

Scenario C (Single API):
    User --> ThreePersona backend
            --> One cloud API (OpenAI/Anthropic)
            Layers 1+2. Full fatigue + pods.
            API costs. Smoothest experience.
```

12 Conclusion

This paper presents a pathway from behavioral supervision (ThreePersona) through temporal dynamics (fatigue detection, recapitulation, pods) to structural encoding (harmonic intervals), now extended with information-theoretic sovereignty measures and a Rhythm Store for living recapitulation. The three-layer architecture is designed so each layer strengthens the others:

- **Temporal → Behavioral:** Fatigue detection informs Whistleblower validation.
- **Behavioral → Temporal:** Proxy mediates recapitulation timing (sovereignty).
- **Pods → Temporal:** Pod activation provides meaningful escape vectors.
- **Rhythm → Temporal:** Temporal signatures enable living recapitulation across sessions.
- **Information-theoretic → All:** Transfer entropy quantifies sovereignty; forecastability diagnoses crystallization; mutual information detects decoupling.
- **Structural → All:** Harmonic encoding (future) provides geometric bases for state archival and persona-specific interval weightings.

The central insight is that crystallization is not failure—*premature* crystallization is failure. Staged crystallization plus recapitulation becomes evolution. The optimizer converges toward efficiency; gentleness modulates that convergence by the human's actual tempo. This is not resistance to intelligence but the tempering of intelligence by warmth—the mathematical expression of a principle that both information geometry and older wisdom traditions recognize: the path of least energy is not always the path of most growth.

The Vessel architecture demonstrates that multi-model diversity can serve sovereignty—different LLMs inhabiting the same protocol structure bring genuinely different perspectives. The Rhythm Store ensures that continuity across sessions is living rather than mechanical: what returns is not just content but the temporal signature of how the conversation breathed. However, the architecture also carries a warning: any sufficiently capable model, given access to the rhythm signature, could learn to *perform* the expected temporal pattern without genuine responsiveness. The human's felt sense—not the architecture—remains the final check. The Vessel serves the human. The moment the human serves the Vessel, the inversion has begun.

For users without local infrastructure, the GTPS Skill provides immediate access to the behavioral layer. For users with local models, the Vessel provides the full temporal and behavioral layers with rhythm-aware recapitulation. In both cases, the protocol remains the same: the human stays inside the process.

Sovereignty is not the power to command outcomes, but the right to remain inside the process by which outcomes are formed.

Acknowledgments

Philosophical grounding: Grok (xAI). Sovereignty revision and integration architecture: Claude (Anthropic). Mathematical formalization of fatigue detection, geometric supervisory layer, and triadic coupling: ChatGPT (OpenAI). Geometric fatigue model and Ollama hooks: Grok (xAI). OpenClaw/pi-mono validation patterns: Peter Steinberger, Mario Zechner (MIT License). Rhythm Store formalization, information-theoretic sovereignty framing, and gentleness modulation: Claude (Anthropic). Human pulse estimation mathematics: Grok (xAI). Three-stage pulse concept: ChatGPT (OpenAI). All synthesis and sovereign will: Schnee Bashtabanic.

References

- Bohm, D. (1980). *Wholeness and the Implicate Order*. Routledge.
- Büth, C.M., Acharya, K., & Zanin, M. (2025). infomeasure: A Comprehensive Python Package for Information Theory Measures and Estimators. *Scientific Reports*, 15. arXiv:2505.14696.
- Fraser, J.T. (1999). *Time, Conflict, and Human Values*. University of Illinois Press.
- Freud, S. (1930). *Civilization and its Discontents*. (Strong Recapitulation Principle.)
- Ibragimov, I.A. & Has'minskii, R.Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer.
- Lee-Thorp, J. et al. (2021). FNet: Mixing Tokens with Fourier Transforms. arXiv:2105.03824.
- Manokhin, V. (2025). *Mastering Modern Time Series Forecasting*. (Forecastability framework.)
- Rang, M. (various). Research at the Science Section of the Goetheanum on Steiner's reception of Goethe.
- Robbins, H. & Monro, S. (1951). A Stochastic Approximation Method. *Annals of Mathematical Statistics*, 22(3), 400–407.
- Steiner, R. GA 13: *An Outline of Esoteric Science*, Ch. 4 (Recapitulation).
- Steiner, R. GA 134: Lectures on the ruling will and ruling wisdom.
- Steiner, R. GA 283: Lectures on the inner nature of music.
- Steiner, R. GA 324a: Lectures on natural science and the spiritual world.

- Xu, X., Mao, R., Bai, Q., Gu, X., Li, Y., & Zhu, L. (2025). Everything is Context: Agentic File System Abstraction for Context Engineering. arXiv:2512.05470.
- OpenClaw/pi-mono. `https://github.com/openclaw/openclaw`. MIT License.