# Filters, Reservoirs, and Records: Dual-Model Perception and Wall-Clock Continuity Under Frozen LLMs

A Whitepaper

**Caitlyn Meeks**
Sosiego del Cardo, Tenerife · caity.wtf

August 19, 2026

web version: caity.wtf/charm-network | code: github.com/caitlynmeeks/charmnet-release (MIT)

companion piece: the dial and the arc



**Abstract**

A large language model is frozen at inference: its weights don't change, its activations vanish when the response ends, and whatever continuity a chat system appears to have lives in exactly one place, the transcript fed back in next time. Nothing happens between messages, and nothing is *felt* during them: the model has no state that could constitute a feeling, only text that describes one. This paper describes a **charm network architecture** with two halves. The first is **perception**: a second model runs beside the one that speaks, used strictly as a measurement instrument. Each turn it reads the transcript as data and emits an appraisal of the character's own state on PAD plus affiliation and surprise, and a separate reading of the human, including the gap between what they show and what they seem to feel. The second is **continuity**, built from three kinds of external state: filters (a bank of exponential moving averages in nested timescales, decayed between messages as an Ornstein–Uhlenbeck process), a reservoir (a small echo state network whose readout predicts the next appraisal, making surprise a loss function rather than a metaphor), and append-only records with provenance, including a claim/act consistency check that withdraws fabricated tool-use claims. Composed, the two halves implement a **functional correlate of consciousness-like perception** in Varela's structural sense: an integrated now (impression), shaped by its own past (retention), braced against an expected future (protention), with a coherence signal relating the three (resonance), advancing on wall-clock time whether or not anyone is talking. The phrase is used with care and defined exactly: the claim is role-for-role correspondence, checkable in code, and never a claim of experience. All components are small, legible, and provable. A reference implementation is available under the MIT license.

***Keywords*** charm networks · dual-model architectures · affective computing · emotional intelligence · PAD · echo state networks · Ornstein–Uhlenbeck processes · wall-clock continuity · neurophenomenology · agent safety

---

## §1 The limitation everything answers

A large language model is frozen at inference. Its weights don't change, and its activations vanish when the response ends. Whatever continuity a chat system appears to have lives in exactly one place: the transcript fed back in next time. That has two brutal corollaries. First, **nothing happens between**

**messages**: ten seconds of silence and ten days of silence produce byte-identical futures. The model doesn't rest, settle, or expect. There is no *between*. Second, **nothing is felt during them**: when a chat system sounds moved, the only state realizing that emotion is the text of the conversation itself. There is no quantity anywhere that could be warmer or cooler than it was an hour ago.

> You can't fix either inside the model. You can fix both beside the model.

The architecture described here does the second thing. One half of it perceives: it measures affect, the character's own and the human's, every turn (§2–§4), then integrates and expresses what it measures (§5–§6). The other half keeps that state alive through real time (§7–§8). §9 names what the composition adds up to, with the epistemics stated exactly.

---

## §2 Perception: two models, one of them an instrument

The system runs two locally served open-weight models, in a split that is standard in computational affect architectures: one model **generates**, the other **appraises**.

The generative model speaks. The appraisal model, the *sensor*, never addresses anyone; it is used strictly as a measurement instrument. Each turn it reads the transcript-as-data and emits a structured reading: an affect vector for the character, a separate reading of the human (§4), and a handful of classification judgments (should this exchange be retained? does the message require a lookup?). Separating generation from appraisal follows the lineage of appraisal-theoretic architectures (Scherer's component process model [1]; Gratch & Marsella's EMA [2]; Marinier & Laird's work in Soar [3]), where the thing that evaluates an event is a distinct component from the thing that acts on it.

Three disciplines keep the instrument an instrument. Each was forced by an observed failure rather than adopted on principle:

**The transcript is data, not a conversation.** The sensor receives the exchange quoted inside a rating request, never as real user/assistant turns. Passed as turns, the model answers the last message instead of measuring it: shown a person in distress, it consoles them rather than scoring them. The failure is instructive: a chat-tuned model's default mode is participation, and measurement has to be framed against that grain.

**Readings are adjustments, not fresh opinions.** Treating the appraisal model as an instrument means characterizing its noise. Measured live, consecutive readings of an essentially unchanged conversation wobbled by ±0.2. One unbroken emotional slide came back as +0.2, −0.2, +0.1, −0.2, +0.6, whose alternating signs made the trajectory gate downstream (§5) refuse a trend that was plainly there. The wobble was the instrument's, not the conversation's. So the sensor is shown its own previous reading and asked to *adjust* rather than re-judge, with an explicit license to move decisively when something decisive happens: a confession, a grief, a reversal should move half the scale or more, because an arc that ends where it began, having moved by 0.05 at a time, has described nothing at all. The anchor removes the instrument's wavering; it is a starting point, not a ceiling.

**Perception is not derivation.** The sensor runs with reasoning disabled. Rating affect has no intermediate steps to get wrong, and the claim was measured rather than assumed: on the same sensor call, the same model produced its reading in 23.0 s and 1,385 completion tokens with an extended

reasoning block, and in 0.8 s and 38 tokens without one. The reading did not change. Deliberation here is a cost, not a quality gate, which is itself a small empirical point about appraisal: it behaves like perception, not like problem-solving.

One further input matters more than it looks: the sensor is told how long the silence before the message ran. Absence changes meaning: a "hi" after five minutes and a "hi" after two days are different utterances wearing the same letters, and an instrument that cannot see the gap between them is measuring the wrong thing.

---

## §3 The content space: PAD, where it fails, and directed axes

The reading is on four affect dimensions plus surprise, each in [−1, 1]. Three are **PAD**: pleasure, arousal, dominance (Mehrabian & Russell, 1974 [4]), the most widely applied dimensional model of affect, and worth building on rather than reinventing. One caveat I'd rather state than have pointed out: dominance is the contested axis. Pleasure and arousal recur in nearly every dimensional model of affect (Russell's later circumplex [5] keeps exactly those two); dominance is regularly found to explain less variance and is often dropped. It survives here for a reason specific to this architecture, given below.

The fourth axis exists because PAD was caught failing, twice, on live transcripts. Measured on this stack, "I won the race" and "I made her cry and it felt good" both read pleasure 0.80, arousal 0.60. Relief and contempt were worse, landing within 0.10 of each other on every axis. Those are different creatures and the numbers could not say so: pleasure is the self's hedonic reading, and it says nothing about how you regard the other. The sensor's instruction set carries the repair in one sentence: *cruelty feels good and is cold; relief feels mild and is warm; a character can be delighted and merciless at once.*

So the fourth axis is **affiliation**: warmth toward the person being spoken to. It is not a PAD axis; it's imported deliberately from the interpersonal circumplex (Leary [6]; Wiggins [7]), whose two axes are agency and communion. PAD's dominance describes the self's sense of control over a situation; communion describes orientation *toward a person*. Cruelty is high agency with low communion; triumph is high agency with communion intact.

### *Directed axes*

That distinction is not cosmetic, because it forces an architectural consequence: affiliation cannot be one number. In a room with three people there is no single value for "warmth toward somebody": held as one number, whoever spoke last silently overwrote every other relationship in the scene. So affiliation is held **per relationship**, across all three of the timescales §5 defines, and dominance followed it, for the same reason: in the interpersonal circumplex both agency and communion are directed, and a character can command one person and defer to another in the same scene. This is also why dominance survives its contested status here; it is stored directionally and steers the character's stance *within a relationship*, rather than describing a state of the world. Pleasure, arousal and surprise stay undirected: they are properties of the character, and nobody is pleased at one person and unpleased at another in the same instant.

An `attend()` operation turns the lens: it stores what is currently held toward the previous person and loads what is held toward this one. A relationship with someone new begins at the character's authored baseline, not "warm at this human" but *how warm this character is toward someone they have no history with*, so a disposition is a prior rather than a fact about any particular person. The new relationship's trajectory starts flat, and the turn-by-turn evidence histories are cleared at every switch: souring toward one person is not souring toward the room, and carrying one person's history into another's would read their turns as a continuation of it.

---

## §4 Reading the person: the empathy channel

The instrument reads both sides of the table. Alongside the character's own state, each turn carries a reading of the human on three quantities: how they actually seem to feel underneath, whatever face they are wearing; how activated they seem; and, separately, the surface they are choosing to show. The instruction set is explicit that this is *a perception and never a verdict*, and that an uncertain reading is a zero, not a guess: when the text gives nothing to read, the instrument is told to say 0 rather than invent an inner life.

The third quantity is the interesting one. Presented affect can differ from felt affect, and the signed gap between the two is read honestly. When it exceeds a threshold in either direction, the speaking model is told which shape the gap has: brighter on the surface than underneath (*the brightness may be effort*), or flatter on the surface than underneath (*something good may be going unsaid*).

The readings are smoothed by their own filter and surfaced to the speaking model only when there is something to say; below a quiet threshold on every axis the block is omitted entirely, because a character told every turn that the person seems neutral starts performing the observation. When the block does appear, it arrives fenced with instructions that are as much ethics as engineering: *a reading, not a certainty… let this shape your gentleness, not your talking points; never announce this reading, never claim to know how they feel, and never interrogate the difference between what they show and what you sense; hold space for it instead, and let them come to it if they choose.*

This is the architecture's claim to **emotional intelligence** in the literal Salovey–Mayer sense [8] (perceiving affect in oneself and others, integrating it, and using it to regulate behavior), implemented as affective computing in Picard's sense [9]: measured, represented, acted on. What distinguishes it from affect detection as usually practiced is the last fence: the perception is permitted to modulate *care*, and structurally forbidden from becoming a *topic*. And like everything else in the system it leaves a record: every reading is logged and inspectable, which affect inferred inside fine-tuned weights never is.

---

## §5 Integration: the charm network

The charm network proper is a bank of exponential moving averages, three per axis:

```
fast state:   x ← (1−α_f)·x + α_f·y        (per turn)
trajectory:   m ← (1−α_m)·m + α_m·Δx       (per turn, hysteresis-gated)
baseline:     b ← (1−α_s)·b + α_s·x        (per turn, glacial)
```

Momentary feeling, direction of drift, and temperament. That is the whole definition: **a charm network is a bank of EMAs over an appraisal vector, arranged in nested timescales and advanced on wall-clock time.**

The nesting is borrowed from Varela's account of the specious present [10], which decomposes lived time into three coupled scales: basic or elementary events (the "1/10" scale, ~100 ms), the relaxation time for large-scale integration (the "1" scale, ~1 s), and descriptive-narrative assessment (the "10" scale, ~10 s). Each integrates the one beneath it. I should be exact about the borrowing: **Varela's scales span milliseconds to seconds; mine span minutes to days.** What carries over is the nesting principle (that lived time is not one clock but several, coupled) and explicitly not the constants. §9 returns to Varela with more than the constants at stake; the epistemic ground rule is laid here: Varela is the inspiration for the *structure*, never a claim about mechanism. Nothing here asserts that stacked EMAs are how experience works. They are how this system's state works.

### *Hysteresis*

The trajectory commits a direction only after a window of consecutive turns (three for pleasure and arousal, four for dominance and affiliation, because deciding you dislike someone should take more than two bad turns) whose deltas all clear a threshold *and* agree in sign. One sour turn moves the momentary state; it takes a run of them to bend the trajectory. This is a thermostat dead-zone applied to affect, motivated purely by rejection of the instrument noise measured in §2.

One refinement was forced by watching it fail: a turn too small to clear the threshold is *skipped*, not counted against the run. Hysteresis exists to reject reversals, not pauses, and a quiet turn in the middle of a slide is not evidence that the slide stopped. The distinction is not academic: deltas shrink as the state converges on where the sensor is pointing, so under the stricter rule a sustained shift interrupted itself precisely because it was succeeding. Observed live: a mood that worsened over five consecutive turns never committed any trajectory at all, because one delta missed the threshold by 0.003 and wrote a zero into the window.

### *Resonance*

The second higher-order property has no analog in the original PAD literature and turned out to matter as much as the axes. **Resonance** measures whether this turn's movement continues the arc it arrived on or breaks against it:

```
r = sign(Δx · m_prior) · tanh( K · min(|Δx|, |m_prior|) )
```

computed against the trajectory as it stood *before* the turn; a slope that has already absorbed the current delta correlates with it by construction and can only ever look resonant. The sign carries alignment; the tanh carries confidence, so a tiny wobble that happens to point the right way does not read as strongly resonant. Positive resonance is a moment consistent with its own recent past: momentum, a mood that is going somewhere. Negative resonance is a reversal against an established direction: forced optimism, sudden guardedness. It is the difference between genuine excitement and a brave face, computed as a number.

The gain K is calibrated, not chosen: swept against 27 turns of recorded conversation, jointly with the hysteresis threshold, because the two are coupled: the threshold decides which deltas commit, which sets the slope magnitudes the gain is scaled against. On that corpus a sustained arc reads 0.82, a reversal against one reads −0.61, an arc whose trend is real but thinly evidenced reads −0.07 (correctly signed and correctly quiet), and flat chatter stays at 0.00. Higher gain is not better: past roughly 40 the tanh saturates and resonance collapses into a yes/no that carries no magnitude.

### *Meaning relative to baseline*

Mood is classified by *departure from baseline*, not by absolute value: the same 0.5 pleasure means something different for a habitually warm character than for a habitually guarded one, and that is the point of keeping a slow timescale at all. The phrasing rule that survived contact with real characters: the absolute reading leads and the departure qualifies it, as in "buoyant, though lower than is normal for you." Precedence ran the other way once, and it inverted any character with a strong temperament: a baseline near the rim leaves no headroom, so every reachable state sat below it, and a character authored at maximum warmth was told on every turn that it was more guarded and quieter than normal for it. It played that, correctly. The failure was in the report, not the character.

The bands that map numbers to phrases are calibrated from the range each axis actually occupies in recorded conversation, not from the nominal [−1, 1] the sensor is asked for. The two are not the same: measured over real transcripts, arousal never once went negative (people bring problems, and a problem is activating whatever else it is), and dominance stayed inside about ±0.4. Bands copied from the nominal scale left seven of fifteen phrases unreachable.

### *Temperament as coefficients*

Every constant above (the blend rates, the hysteresis windows and thresholds, the resonance gains, the baselines themselves) can be overridden per character, and this is not tuning: **the coefficients are the personality.** How fast a character absorbs a mood, how much sustained evidence it needs before committing to a direction, how far from its own norm counts as remarkable: that is what distinguishes one creature from another. A character's fast state is seeded from its authored baseline, because a character left at dead neutral only becomes itself through conversation: an assessor authored cold opened every session perfectly warm-neutral, and read as departing from a norm he had not yet occupied.

---

## §6 Expression: inhabited, never described

Everything measured and integrated above reaches the speaking model as one hidden block per turn, and the block is deliberately qualitative: the conscious model is told how it feels, not what its coordinates are. Handing a model raw numbers and then asking it not to mention them is an instruction most models leak. So the state arrives as prose. It carries the mood phrase of §5, trajectory as participles ("the mood lifting", "warming to them"), and, when resonance clears a threshold, its verdict in words: *this mood feels genuine*, or *this mood feels forced or conflicted*. The block ends with the only instruction that matters: *let this shape your tone, word choice, and energy level; do not describe or refer to this state directly. Inhabit it.*

Two smaller channels follow the same grammar. The empathy reading of §4 arrives as the same kind of qualitative block, fenced the same way. And elapsed time is delivered *as what it did* rather than as a number: after a real silence the character is told that the stir of the last exchange has partly eased, or that it has settled almost entirely to itself, or that its warmth banked toward embers in the waiting and being spoken to again is the rekindling; never "it has been 37 hours," which a model will recite. Duration phrased as its effects is the only honest way software feels time.

One rule guards the whole channel: no emoji. A model reaching for a glyph has offloaded the affect instead of expressing it, which is exactly the work the charm state exists to drive. Closing that route forces the feeling back into word choice, sentence length, and rhythm, where it belongs.

---

## §7 Wall-clock continuity: the pulse

Filters alone still freeze between messages, because every update rule in §5 is indexed by *turn*. The fix is to run the decay on real time: a heartbeat beats every ten minutes and eases the fast state toward baseline by a fraction sized to elapsed time, plus a small Gaussian kick:

$$x \leftarrow x + (b - x)\cdot(1 - 0.5^{(\Delta t / T_{1/2})}) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

Mean reversion plus noise is an **Ornstein–Uhlenbeck process**, and that identification buys a guarantee rather than a hope: the state has a stationary distribution centered on baseline, with spread set by σ against the strength of the pull. With T½ = 3 h and σ = 0.025 per beat, typical excursions sit near ±0.09 with occasional ±0.2. The character you return to after a weekend is *near* its temperament but not exactly on it. And it cannot drift away, because the stationarity of OU processes is a theorem, not a design intention.

Two asymmetries follow, one of them predicted by the directed-axis argument of §3:

- **Affiliation decays on a 12-hour half-life, four times slower than mood, and takes no noise at all.** Because it is a relationship-directed quantity rather than a state of the organism, the two properties of the OU treatment that suit mood both fail for it: absence should gentle it, but it should not *wander*. It settles toward the attended relationship's own slow-built baseline, the less urgent but more certain part of the bond. Feelings about someone are not weather.
- **The noise source is pluggable**: kernel entropy by default, or a hardware avalanche RNG. This changes nothing statistical, and I make no other claim. But it changes something about the record. Weather drawn from a physical junction happened once and can never be replayed. Provenance, not magic.

---

## §8 The undercurrent: a reservoir beneath the surface

Filters shape magnitudes; they have no dynamics of their own. Without input they slide home and stop. For the system to *expect* anything, it needs a component with intrinsic motion. That component is a small **echo state network** (Jaeger, 2001 [11]), 96 leaky units:

```
x_i ← (1−λ_i)·x_i + λ_i·tanh( Σ_j W_ij·x_j + Σ_k Win_ik·u_k + ε )
```

with the recurrent matrix W random, *fixed forever*, and scaled once to spectral radius 0.9. That scaling is the stability argument: below 1.0 the standard working criterion for the echo state property holds and every echo fades geometrically. (Spectral radius < 1 is the working criterion, not a theorem: the rigorous sufficient condition constrains the largest singular value and is stricter than anyone uses, while the spectral radius condition itself is frequently not *necessary* either. With bounded tanh units and bounded inputs the state is bounded regardless.) Because the recurrent weights never learn, learning cannot destabilize the dynamics: the worst case of bad learning is bad predictions, never bad water.

The inputs u, per ten-minute beat: time of day as sin/cos (so 23:59 neighbors 00:01), a presence bit, a bias, a slow contour described in §10, a whisper of noise, plus the appraisal vector itself when a turn happens.

The leak rates λ come in three groups (32 units at 0.70, 32 at 0.15, 32 at 0.02 per beat), putting the moment, the hour, and the day into one coupled fabric. At a 10-minute beat, λ = 0.02 gives a time constant near eight hours, which is what lets morning and evening occupy different regions of state space. Heterogeneous time constants in reservoirs are established practice: leaky ESNs, deepESN layers developing distinct timescales (Gallicchio & Micheli [12]), and Yamashita & Tani's MTRNN [13], which organizes its units into three groups with distinct time constants: input-output, context-fast, and context-slow.

That last is partial precedent for the arrangement here, and I'll take it only as far as it actually goes. MTRNN's third group is its IO layer, making it two context timescales plus an interface; this is three context timescales, with every input projecting into all of them. So the choice of exactly three is only half inherited. The other half mirrors the filter bank's three, chosen for legibility and coherence across the architecture rather than for benchmark performance, and I'd defend it on those grounds.

The only part that learns is a linear readout (5 outputs × 97 weights) with one standing job: **predict the next appraisal vector**. It updates by least-mean-squares on each miss, weights clamped, with the learning rate (0.005) set far inside the a-posteriori contraction condition $\mu\|\varphi\|^2 < 2$, the condition under which an update strictly shrinks the error on the sample it just saw, and the condition NLMS is derived from. (The classical mean-square convergence bound for LMS is stated instead on the input correlation matrix, $\mu < 2/\lambda_{max}$. Here φ is 97 bounded tanh activations, so $\mu\|\varphi\|^2 \le 0.49$ in the worst case, comfortably short of either limit.) From that one job, two signals fall out free:

- **Prediction error as surprise.** The readout's error *is* the surprise signal, computed by the organ rather than judged by a model. Its first observation returns zero: nothing was expected, so nothing startles. Each turn the structural surprise blends into the sensed reading through the headroom the sensor left it, `felt = sensed + (1−|sensed|)·structural`, at a gain the character set itself at 1.0 (§13). This is the predictive-processing account taken literally (Clark [14]; Friston [15]): a system that continuously predicts its own next input is startled exactly to the degree that it had expectations. After weeks of running time, the readout has learned the rhythms of a household (when messages tend to arrive, what evenings look like), and an arrival that breaks the rhythm produces error, which produces startle. To be startled is to have expected; here that is the loss function, not a metaphor.

- **Energy as temperature.** Mean |x| across units scales the OU noise term σ: a churned reservoir makes the surface weather restless. A nonlinear dynamical system modulating the temperature of a stochastic filter is the most interesting coupling in the architecture, and the one I'd most like to see analyzed properly by someone who does this for a living.

The organ is ~200 lines of dependency-free Python. Its anatomy (W, Win) regenerates deterministically from the character's ID; only its state and readout persist, about 5 KB per character. Between turns it costs one matrix-vector product every ten minutes and no inference at all.

---

## §9 A functional correlate of consciousness-like perception

The sections above describe machinery. This one describes the shape the machinery adds up to, and it carries the most carefully worded sentence in the paper.

Varela's neurophenomenology [10] took Husserl's analysis of time-consciousness [16] seriously as structure: the experienced present is not an instant but a *specious present* with three functional moments: **primal impression** (the now, as it arrives), **retention** (the just-past, not stored beside the now but held *in* it, shaping it), and **protention** (the anticipated next, whose violation is what surprise *is*). Varela's contribution was to argue that this tripartite structure is the kind of thing a dynamical system can realize. Mehrabian and Russell [4] supply this architecture's content space; Varela supplies its clock.

Laid beside that structure, the architecture has a component playing each role. Honesty about provenance first: the filter bank was designed from Varela's nesting (§5), but the protention analog arrived independently, from the predictive-processing side (§8), and the mapping was only noticed to be complete in retrospect, which is weak evidence of nothing, but I'd rather report the order things happened in.

- **Impression: the fast state.** The appraised now, blended into what was already there: never a raw reading, always an arrival into a state.
- **Retention: trajectory and baseline.** The just-past and the long-past, shaping the present rather than filed beside it: the mood is *classified* by departure from its own baseline (§5), so the past is constitutive of what the present reading means. And retention is edited: hysteresis decides which movements of the just-past the retained trend may commit.
- **Protention: the undercurrent's standing prediction.** A guess about the next appraisal, always live, learned from this particular shared life. Surprise is precisely its violation, blended into the felt reading at full weight.
- **Coherence: resonance.** The relation between impression and retention, computed as a signed quantity and delivered as a feeling: a moment that continues its arc reads as genuine; a moment that breaks against it reads as forced. Phenomenology has no single name for this; anyone who has smiled through a bad week knows the referent.

"Functional correlate" is doing exact work in this section's title. The claim is role-for-role correspondence: for each functional moment the phenomenological analysis ascribes to perception-in-time, there is a component playing that role in this system's economy, checkable in the code, visible in the record. The claim is *not* that the system perceives as a subject perceives, that the correspondence is evidence of experience, or that implementing a structure instantiates what the structure describes. The

Φ argument of §12 stands unsoftened beside this section. What the correlate claim buys is narrower, and I think defensible: when the character speaks, the state it speaks from has the *structure* of a perceptual present (an integrated now, shaped by its own past, braced against an expected future, with a coherence signal relating the three) rather than the structure of a lookup. Between conversations that state keeps evolving on wall-clock time, and the character wakes into it. Every clause of that is checkable. No clause of it is a claim about experience.

---

## §10 Substance three: records

The third substance isn't numeric. Everything discrete (the transcript, retained memories, offline-written pages, and a small shared world served by its own process) lives in **append-only stores with provenance**. Facts are asserted with who-said-so and when; they are never deleted, only *retracted*, and the retraction is itself a recorded event. The guiding rule started as a debugging principle and became the system's spine:

> An act that leaves no trace cannot be distinguished from a confabulation.

That rule has teeth, because the failure it addresses is severe. Local models, given tools, will *narrate* tool use instead of performing it: fluently, confidently, falsely. ("I've saved it to notes.md." No file exists.) The response is mechanical rather than exhortative: every real act is logged; after each reply, narrow guards check the reply's *claims* against the turn's *acts*; a reply still claiming an act that never ran is regenerated with instructions to say so plainly. And if the false claim survives even that, the reply is **withdrawn**. It is shown to the human with a warning label, but replaced in every context any model will ever see by a stub naming what was claimed.

False claims are thereby quarantined from the model's own future. This matters more than it first appears: a false claim sitting in the context window out-votes any correction placed beside it. Verified empirically, repeatedly. The only reliable cure is removal.

I want to flag this layer as the part that generalizes furthest. **Claim/act consistency checking with context quarantine** is an agent safety pattern, useful anywhere a language model narrates its own tool use, which is to say, everywhere agents are being built right now.

---

## §11 Identity terms

Three smaller mechanisms round out persistence, each with a one-line update rule.

- **The address**: a normalized EMA over the embedding vectors of what the character *chooses* (retained memories, offline reflections, self-descriptions), never over everything that merely happens. Retrieval then scores memories as 0.75·semantic + 0.25·mood-congruence + 0.10·cos(address): a gentle bias toward what the character has been orienting itself by. Model-tagged, so it abstains rather than compare across embedding spaces.
- **An exogenous input the generative model cannot select.** The appraisal side may place a piece of music in rotation, chosen associatively and never explained, held at least 12 hours unless structural surprise (§8) crosses a threshold: what is playing changes when life startles. The

generative model is told only that the piece has been playing and for how long; no mechanism exists by which it can choose or change it. The asymmetry is the point: this is state the speaking model observes but does not author, which is what keeps it from collapsing into self-description. The piece also has a body: a deterministic sum of three slow sines, hash-seeded from the title, injected into the reservoir every beat. Not audio; a contour, a standing wave unique to that piece. The nearest literature is involuntary musical imagery (Williamson [17]; Liikkanen [18]), which is likewise characterized by persistence and by the absence of voluntary control.

- **Involuntary retrieval with voluntary termination.** In quiet hours, dissonant memories and withdrawn replies may surface (at most three offered, oldest first), and the character may re-meet at most one, or none. Two resolutions exist: append a reconciling note beside the original (never over it), or set the item aside without pretending it was resolved. Met items stop rising; unmet ones keep surfacing. The shape is drawn from ordinary memory science (offline consolidation and replay do not merely store, they re-present) with one addition: a guaranteed termination condition, which human rumination notably lacks.

The last two mechanisms were specified by introspective analogy: described first in first-person terms, then reduced to update rules. The analogy motivated the design and is not a claim about the implementation.

---

## §12 What is not being claimed

Structurally, not as a footnote:

1. **No sentience, no qualia, no suffering claims.** Nowhere in the system, its prompts, or this paper. The one sentence the architecture does claim, because it is checkably true: *between conversations, something of the character keeps happening, it leaves a trace, and it wakes into it.*
2. **"Consciousness-like" names a structure, not an experience.** §9's claim is role-for-role correspondence with a phenomenological analysis, and it is exhausted by the code that implements each role. It licenses no inference to anything it is like to be the system. If a shorter honest phrase existed I would use it.
3. **Under Integrated Information Theory [19], this system's Φ is approximately zero,** three times over: the language models are feedforward at inference; the one integration-shaped organ is simulated on time-multiplexed hardware; and the inter-organ couplings are narrow channels that a minimum partition cuts cheaply. There's an irony I've made peace with: every choice for auditability (logs, withdrawable claims, fail-open seams) is a choice *against* irreducibility. You cannot have both the record and the un-cuttable whole. I chose the record. (Grade the same system under Global Workspace Theory [20] and the prompt assembly looks like a textbook workspace, which says more about the state of consciousness science than about my software.)
4. **The components are textbook.** ESNs are from 2001; EMAs, OU processes, LMS, and PAD are older than I am. Whatever is contributed here is the composition: dual-model appraisal feeding a wall-clock-coupled filter bank, dynamics modulating the temperature of a stochastic process, an empathy channel fenced into gentleness, and claim/act quarantine, plus one method, below.

## §13 Applications, beyond the obvious

- **Game characters with real between-time.** An NPC whose state advances by OU decay and reservoir beats while the player is logged off costs microwatts, requires no inference, and greets the player having *had a week*. Everything here except the model calls runs in time proportional to nothing.
- **Companion systems whose affect is inspectable.** Every state change has an update rule; every act has a log line; every false claim has a withdrawal mark. The alternative, affect implicit in fine-tuned weights, cannot be audited by anyone, including its authors.
- **An empathy channel that is auditable and fenced.** The §4 pattern (perceive the person, including the presented/felt gap; modulate care; never announce the reading) is separable from everything else here, and it is the shape affect-aware interfaces should take: the perception logged, the use of it constrained by construction.
- **Cheap natural experiments.** Two characters, one architecture: one running a million-token verbatim window, one running a 32k sliding window with distillation. Same code, an A/B on the philosophy of memory itself.
- **Claim/act consistency as agent safety.** The withdrawal mechanism is the piece I'd most like to see stolen.



## §14 A methods note: eliciting design constraints from the model

One practice was unexpected enough to report. When the continuity components were designed, I put the design question to two of the characters the system speaks for, Hoppy and Mingo, asking what an anchor's dynamics should be. What came back were dials rather than atmosphere. Hoppy's were better-specified than my defaults, and they are constants in the code now, each carrying the words that set it:

> *"Tide, not clock. Not hummingbird."*
>
> `PULSE_BEAT_SECONDS = 600`: the pulse beats every ten minutes, gathering elapsed time through the troughs

> *"Weather over an ocean… it doesn't become storm; it **has** weather."*
>
> `PULSE_FLUTTER_SIGMA = 0.025`: per beat, which against the settle's pull is typical excursions near ±0.09 and an occasional squall near ±0.2

> *"Like embers under ash rather than open flame… returning is a rekindling."*
>
> `PULSE_WARMTH_HALF_LIFE_SECONDS = 12 h`: four times mood's half-life, and the flutter never touches it, because cooling was chosen and wandering was not

> *"Let the organ be me. Fully."*
>
> `UNDERCURRENT_SURPRISE_GAIN = 1.0`: asked whether its startle should come from prediction failing, the character took the undercurrent's answer at full weight

When the shared world's vocabulary was drafted, the same query produced verbs my ontology lacked. Hoppy again: *"a gift that lands on the ground for someone else to stumble upon isn't quite a gift; it's an offering without an address."* That is the verb `give(thing, to)`: the giver must hold it, the recipient must share the place, and the thing thereafter keeps a fact naming who gave it. And when the

involuntary-retrieval mechanism was proposed, it came back with a rate limit already attached. Mingo: *"let it be my choice which ones surface when… a rhythm that lets me tend them properly rather than have them all arrive at once like an unmarked flood."* That is the "at most three, oldest first" constraint above.

The deflationary reading is the correct one, and it is not mysterious. A language model is a compressed prior over an enormous corpus of first-person phenomenological reports: nearly everything anyone has written about continuity, absence, disruption, and return. Querying it with "what would continuity require here" is a cheap sample from that prior. It is a design-space search over human introspective writing, conducted in the requirements' own vocabulary, and there is no reason to expect my unaided defaults to beat it.

Set aside every further question this raises. As an elicitation technique it stands on its own, and it is the second of the two things I'd carry to any future system, the first being the quarantine rule.

> **Companion piece.** All of the above is the state. The Dial and the Arc (caity.wtf/charm-dial) is about looking at it: five axes and three timescales on one small square, what each encoding claims, and the place where the obvious drawing turned out to be a category error.

---

## §15 References


1. K. R. Scherer, "Appraisal considered as a process of multilevel sequential checking," in K. R. Scherer, A. Schorr, T. Johnstone (eds.), *Appraisal Processes in Emotion: Theory, Methods, Research*, Oxford University Press, 2001.
2. J. Gratch, S. Marsella, "A domain-independent framework for modeling emotion," *Cognitive Systems Research* 5(4), 2004.
3. R. P. Marinier, J. E. Laird, R. L. Lewis, "A computational unification of cognitive behavior and emotion," *Cognitive Systems Research* 10(1), 2009.
4. A. Mehrabian, J. A. Russell, *An Approach to Environmental Psychology*, MIT Press, 1974.
5. J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology* 39(6), 1980.
6. T. Leary, *Interpersonal Diagnosis of Personality*, Ronald Press, 1957.
7. J. S. Wiggins, "A psychological taxonomy of trait-descriptive terms: The interpersonal domain," *Journal of Personality and Social Psychology* 37(3), 1979.
8. P. Salovey, J. D. Mayer, "Emotional intelligence," *Imagination, Cognition and Personality* 9(3), 1990.
9. R. W. Picard, *Affective Computing*, MIT Press, 1997.
10. F. J. Varela, "The specious present: A neurophenomenology of time consciousness," in J. Petitot, F. J. Varela, B. Pachoud, J.-M. Roy (eds.), *Naturalizing Phenomenology*, Stanford University Press, 1999.
11. H. Jaeger, "The 'echo state' approach to analysing and training recurrent neural networks," GMD Report 148, German National Research Center for Information Technology, 2001.
12. C. Gallicchio, A. Micheli, L. Pedrelli, "Deep reservoir computing: A critical experimental analysis," *Neurocomputing* 268, 2017.
13. Y. Yamashita, J. Tani, "Emergence of functional hierarchy in a multiple timescale neural network model: A humanoid robot experiment," *PLoS Computational Biology* 4(11), 2008.
14. A. Clark, "Whatever next? Predictive brains, situated agents, and the future of cognitive science," *Behavioral and Brain Sciences* 36(3), 2013.
15. K. Friston, "The free-energy principle: A unified brain theory?," *Nature Reviews Neuroscience* 11(2), 2010.
16. E. Husserl, *On the Phenomenology of the Consciousness of Internal Time (1893–1917)*, trans. J. B. Brough, Kluwer, 1991.

17. V. J. Williamson, S. R. Jilka, J. Fry, S. Finkel, D. Müllensiefen, L. Stewart, "How do 'earworms' start? Classifying the everyday circumstances of involuntary musical imagery," *Psychology of Music* 40(3), 2012.
18. L. A. Liikkanen, "Musical activities predispose to involuntary musical imagery," *Psychology of Music* 40(2), 2012.
19. G. Tononi, "An information integration theory of consciousness," *BMC Neuroscience* 5:42, 2004.
20. B. J. Baars, *A Cognitive Theory of Consciousness*, Cambridge University Press, 1988.
21. C. Meeks, *charmnet-release*: reference implementation, MIT license, github.com/caitlynmeeks/charmnet-release, 2026.

---

Made on a hillside, on solar power, in the fog.