# Intentional Fragility: Bayesian Context Decay and Semantic Rot in Local-First LLM Orchestration

Mazze LeCzzare Frazer
Independent Researcher
[contact]

Draft — May 30, 2026

**Abstract**

Prevailing approaches to context management for large language model (LLM) applications optimize for retention and graceful degradation: context is treated as an asset to be preserved. We argue this framing is incomplete. A context window that forgets nothing eventually carries more noise than signal, because stale and low-utility information competes for finite attention against information that currently matters. We present *Context Synapse*, a local-first context orchestration engine built on the inverse principle, which we term *intentional fragility*: context should decay under disuse, exhibit visible semantic rot when it drifts from the active task, and become cheaper to discard when it repeatedly fails to deliver predicted utility. We formalize three mechanisms — a passive exponential decay term, an active semantic rot term bounded by a hyperbolic tangent, and a Bayesian prediction-error feedback loop in which each context unit holds a mutable Beta-distributed prior over its own utility. We show that prediction error — the discrepancy between a context unit's predicted and observed utility — provides a principled, self-correcting signal that existing Bayesian orchestration frameworks do not exploit, and that anchor ("lighthouse") protections derived from this prior are self-correcting rather than permanently granted. We further describe a fault-injection calibration protocol in which the system stress-tests its own coupling topology. We position this work relative to recent Bayesian multi-LLM orchestration [1] and argue that the consent and correctability properties of the design make it suitable for deployment with neurodivergent and clinically-adjacent user populations where inference-as-diagnosis is a material risk.

## 1 Introduction

Context management is the central engineering problem of applied LLM systems. Retrieval-augmented generation, long-context architectures, and memory systems all share a common implicit objective: maximize the retention of potentially relevant information subject to a context-length budget. The dominant design philosophy, reflected across major tooling ecosystems, treats forgetting as a failure mode to be minimized.

We take the opposite position. The scarce resource in an LLM interaction is not storage; it is *attention* — the finite budget of tokens that can meaningfully condition a generation. Within that budget, every retained unit of stale or low-utility context displaces a unit that currently matters. Under this framing the relevant question is not "how much can we retain?" but "what has earned its place, and what has quietly become noise while continuing to present as signal?"

This reframing yields a design principle we call **intentional fragility**: context should cost something to keep. Specifically, a unit of context should (i) *decay* passively under disuse, (ii) *rot*

actively and visibly when its semantic content drifts away from the active task, and (iii) become *cheaper to discard* when it repeatedly overestimates its own usefulness.

The third property is the central technical contribution of this paper. We observe that most systems track *what happened* (recency, access frequency, success/failure of an interaction) but never track *what the system expected to happen*. The discrepancy between the two — prediction error — is the learning signal of every constructivist account of intelligence, biological or artificial, yet it has no architectural home in current context-management systems. We give it one.

**Contributions.**

1. A formalization of intentional fragility as three composable terms: passive decay, semantic rot, and prediction-error amplification (Section 3).
2. A Bayesian prediction-error feedback loop in which each context unit holds a mutable Beta prior over its utility, updated by observed outcomes, such that overconfident-but-useless context self-prunes (Section 4).
3. A correctability result for anchor protection: "lighthouse" floors derived from the prior are self-correcting, eliminating the permanent misdesignation failure mode of static floors (Section 5).
4. A fault-injection calibration protocol by which the system measures its own coupling topology and recommends its own decay parameters (Section 6).
5. A consent-preserving treatment of affect-conditioned context that keeps inference strictly provisional and out of the autonomy path, motivated by deployment with neurodivergent and clinically-adjacent populations (Section 7).

## 2 Related Work

**Bayesian LLM orchestration.** Amin [1] proposes treating LLMs as approximate likelihood models rather than classifiers, eliciting likelihoods via contrastive prompting and updating beliefs with Bayes' rule under explicit priors as evidence arrives. This establishes sequential Bayesian belief updating at the orchestration *control* layer. Our work is structurally aligned: we also maintain and update beliefs under explicit priors. We differ in object and objective. Amin updates beliefs over *task states* to select *actions* under asymmetric error costs; we update beliefs over the *utility of context units* to decide what to *retain or discard*. To our knowledge, no prior Bayesian orchestration framework treats the prediction error of a context unit's utility as a decay-modulating signal, nor addresses the correctability of anchor protections.

**Predictive processing.** The decay–rot–error decomposition draws on predictive-processing accounts of cognition, in which perception and action are driven by the minimization of prediction error against a generative model. We do not claim a neuroscientific result; we borrow the *architecture* of error-driven belief revision as an engineering pattern, and we are explicit about that boundary.

**Context engineering as lifecycle.** Lifecycle approaches to context (generate, evaluate, distribute, observe) treat context as a versioned artifact managed analogously to source code. This is complementary to our approach: lifecycle methods govern how context is *authored and shipped*; intentional fragility governs how context *behaves at runtime* once in play. The two compose; they do not compete.

## 3 Mechanisms of Intentional Fragility

We model the active context as a set of *synapses* $s \in S$, each a unit of context (a file region, a decision record, a prior conversation turn). Each synapse carries a time-varying weight $W(s,t) \in [0,1]$ that determines its priority for inclusion in prompt assembly.

### 3.1 Passive Decay

Under disuse, a synapse's weight decays exponentially:

$$W_{\text{decay}}(s,t) = \bar{\mu}(s)\, e^{-\lambda(s,t)\,\Delta t}\, U(s,t), \tag{1}$$

where $\Delta t$ is the elapsed time since the last successful interaction with $s$, $U(s,t)$ is a recency-weighted utility score, and $\bar{\mu}(s)$ is the *base utility* of the synapse. Crucially, $\bar{\mu}(s)$ is not a fixed constant: it is the mean of a mutable Bayesian prior, developed in Section 4. The decay constant $\lambda(s,t)$ is itself dynamic:

$$\lambda(s,t) = \lambda_{\text{base}} \big(1 - c(s)\big)\, \rho(s)\, \big(1 + e(s) \cdot \gamma\big), \tag{2}$$

where $c(s) \in [0,1]$ is a *connectivity factor* (synapses embedded in a dense relational structure decay more slowly), $\rho(s)$ is a rot multiplier (Section 3.2), $e(s)$ is the prediction error (Section 4), and $\gamma$ is the error-decay amplifier. The final term is novel and is the mechanism by which overconfident-but-useless context is preferentially pruned.

### 3.2 Semantic Rot

Decay handles disuse. *Rot* handles drift: a synapse may be accessed frequently yet have become semantically irrelevant to the active task. We anchor the active task with a designated *lighthouse* synapse and measure each synapse's semantic distance from it:

$$\text{Rot}(s) = D(s,\ell)\, \tanh\!\left(\frac{T_{\text{drift}}(s)}{T_{\text{thr}}}\right) V(s), \tag{3}$$

where $D(s,\ell) \in [0,1]$ is the semantic distance from synapse $s$ to the lighthouse $\ell$, $T_{\text{drift}}(s)$ is the time spent drifting, $T_{\text{thr}}$ is a threshold, and $V(s)$ is a velocity amplifier. The tanh envelope is deliberate: it produces a gradual build rather than a cliff edge, so rot accrues as a warning gradient rather than a sudden cutoff. At $T_{\text{drift}} = T_{\text{thr}}$, $\tanh \approx 0.76$ (warning); at $2T_{\text{thr}}$, $\tanh \approx 0.96$ (cauterization imminent). Semantic distance $D$ admits a complexity ladder: structural name overlap (ship now), TF–IDF cosine similarity (ship soon), and local embedding similarity (MiniLM via on-device inference, future), allowing the same formula to be deployed before an embedding model is available.

## 4 The Prediction-Error Circuit

The base utility $\bar{\mu}(s)$ in Eq. 1 is the mean of a **Beta-distributed prior** over the synapse's utility:

$$p(s) \sim \text{Beta}(\alpha_s, \beta_s), \qquad \bar{\mu}(s) = \frac{\alpha_s}{\alpha_s + \beta_s}. \tag{4}$$

The Beta distribution is chosen because it is bounded on $[0,1]$ (matching the utility domain), is the conjugate prior for Bernoulli-distributed success/failure observations (making updates closed-form and cheap), and exposes its evidence weight directly as $\alpha_s + \beta_s$.

### 4.1 Forward and Backward Passes

The circuit alternates two passes. The **forward pass** generates a prediction for each synapse from its current prior, modulated by relational topology:

$$\hat{u}(s) = \bar{\mu}(s)\,\big(1 + b(s) \cdot \kappa\big), \tag{5}$$

where $b(s)$ is a topological bias term (a weighted average of upstream synapse priors) and $\kappa$ is a small coupling coefficient. The **backward pass** records the observed utility $o(s) \in [0, 1]$ of an interaction (mapped from interaction events: a commit scores 1.0, a file save 0.9, a build failure 0.2, attention leaving the window 0.0) and performs the conjugate update:

$$\alpha_s \leftarrow \alpha_s + \eta\, o(s), \qquad \beta_s \leftarrow \beta_s + \eta\,\big(1 - o(s)\big), \tag{6}$$

with learning rate $\eta(\tau) = \eta_0/(1 + \tau\,\delta)$ decaying in the pass count $\tau$. The decaying rate implements a deliberate fragility gradient: early priors are cheap to overwrite; mature priors resist noise.

### 4.2 Prediction Error as a First-Class Signal

The prediction error of a synapse is

$$e(s) = \big|\, \hat{u}(s) - o(s) \,\big|. \tag{7}$$

This quantity feeds directly into the decay constant via the amplifier term in Eq. 2. The consequence is the paper's central behavioral claim: *a synapse that consistently predicts high utility for itself and then fails to deliver decays faster than a synapse that modestly predicts low utility and delivers.* Overconfident, useless context is the worst case for an attention budget — it both occupies space and misrepresents its value — and the error amplifier targets it specifically. Crucially, error does not move a neighbor's mean: when error propagates across a relational edge, it widens the neighbor's *uncertainty* (increments $\beta$) without penalizing its mean. Adjacent synapses become less confident, not less useful; they are invited to gather more evidence. This preserves the relational structure as a network of *confidence coupling* rather than blame propagation.

## 5 Self-Correcting Anchors

Anchor protection is the standard mechanism for preventing critical context from being discarded: a designated unit receives a weight floor below which it cannot fall. The naive implementation grants a static floor: $\text{floor}(s) = c_0$ if $s$ is an anchor, else 0.

This static floor reintroduces the failure mode the system is designed to avoid. If the anchor designation is *wrong* — the unit was important once but is no longer — the static floor prevents decay from ever correcting it. The error is permanent absent manual intervention. We replace the static floor with a *prior-derived* floor:

$$\text{floor}(s) = \bar{\mu}(s) \cdot c_0. \tag{8}$$

An anchor with a well-supported prior holds a high floor; an anchor whose prior erodes under repeated prediction failures loses its floor in proportion. The anchor remains the anchor — its high starting prior is preserved — but the floor is *earned through demonstrated accuracy* rather than granted in perpetuity. Misdesignation becomes self-correcting through the same prediction-error loop that governs ordinary synapses. To prevent the opposite pathology — an anchor ossifying because its prior accumulates so much evidence that it becomes effectively immovable — we cap the evidence weight $\alpha_s + \beta_s$; beyond the cap, updates are gated behind an explicit, user-initiated reset rather than applied automatically.

## 6 Fault-Injection Calibration

Intentional fragility introduces a meta-question: is the system fragile in the *right* way? A circuit whose synapses are too tightly coupled will cascade a single prediction error across the whole network; one whose synapses are too isolated cannot exploit relational structure at all. We make the system test itself.

The **fault-injection protocol** introduces a controlled, artificial prediction error of severity $\sigma$ into a target synapse and performs a breadth-first traversal of the relational graph, measuring how many hops the fault propagates before attenuating below a meaningful-bleed threshold. Across a full suite over all synapses at two severity levels, the protocol computes a *coupling index* (fraction of injections producing pathological cascades) and an *isolation index* (fraction producing no propagation), and from these recommends values for the rot and cauterization parameters. The protocol runs on a forked snapshot by default, leaving live state untouched; live mutation is an explicit opt-in for deliberate adversarial-hardening sessions. The healthy regime is a propagation depth of one to three hops with graceful attenuation. The system thus dogfoods its own fragility: it does not assume its coupling is correct, it measures it.

## 7 Consent, Affect, and Inference-as-Diagnosis

Context Synapse is intended for deployment with neurodivergent and clinically-adjacent user populations, where a class of ethical risk arises that generic context systems can ignore. If the system infers a user's affective or cognitive state and then *acts* on that inference — silently re-anchoring the active task, or switching interaction modes — it has crossed from detection into autonomous intervention on the basis of an unconfirmed inference about a person. We treat this as a hard boundary and enforce a three-layer separation: *detection* (recognizing a state) and *inference* (forming a provisional belief about it) are permitted and treated as literacy; *autonomy* (acting on the inference without permission) is not.

Architecturally, this means affect signals are computed *asynchronously* and surface as *available context*, never as a trigger. Anchor changes and interaction-mode switches occur only on confirmed user choice. We additionally enforce an epistemic hierarchy in which user self-report is authoritative and model inference is strictly provisional, and we flag a timing side-channel (inferring affect state from the latency of an affect-conditioned pass) as a concrete attack on this boundary, mitigated by fixed-interval scheduling and jitter. The result is a system that can see and name a state and make what it sees available, while leaving the decision with the person. We argue this *correctability* — of anchors, of priors, and of inferences — is the property that distinguishes a deployable design from a paternalistic one.

## 8 Discussion and Limitations

**Validity boundary of the affect proxies.** The deployment substitutes laboratory affect sensors (EEG, eye tracking) with consumer proxies (cursor scanpath, scroll cadence, webcam pupil estimation) and semantic embeddings as a substitute for normed affective image stimuli. This loses population-level normative validation; the proxies are suggestive, not measured ground truth. The consent architecture of Section 7 is partly a response to this: because the inference is provisional and never acted on autonomously, the weakness of the proxy is bounded in its consequences.

**No empirical evaluation yet.** This paper presents an architecture and its formal properties, not a user study. The behavioral claims — that the error amplifier preferentially prunes overconfident-

useless context, that prior-derived floors self-correct misdesignation — are stated as design consequences and are testable; we have not yet run the longitudinal study that would confirm them in deployment. We regard the fault-injection protocol as a necessary but not sufficient internal validity check.

**Single-writer assumption.** The current design assumes a single writer to the synapse store within a session. Concurrent multi-agent writes would require a locking discipline we have specified but not yet hardened.

## 9 Conclusion

We have argued that the right objective for context management is not retention but selective, accountable forgetting, and we have given that objective an architecture. Intentional fragility decomposes into passive decay, active semantic rot, and a Bayesian prediction-error loop in which each context unit holds a revisable belief about its own utility. The prediction-error term gives the system the one capability that recency- and frequency-based methods structurally lack: the ability to learn from surprise. Anchor protections built on these priors are self-correcting rather than permanent, and the system calibrates its own coupling through fault injection. We hold strong beliefs weakly: the lighthouse keeps the light on, but it earns the floor it stands on.

## References

[1] D. Amin. Bayesian Orchestration of Multi-LLM Agents for Cost-Aware Sequential Decision-Making. *arXiv preprint arXiv:2601.01522*, 2026.