

Reflex Attention in Stratified Agent Architectures

A Short-Horizon Deviation Flag for Non-Persistent Anomaly Detection

Stephen A. Putman

putmanmodel@pm.me

Version 0.2 — February 2026

Abstract

Stratified agent architectures separate transient variation from structurally relevant state through memory tiering and governed deviation routing. However, systems that operate under constraint-accessible inputs also require a short-horizon salience mechanism that can elevate attentional weighting without implying persistence, baseline update, or governance escalation. This paper specifies a Reflex Attention layer as a non-persistent anomaly flag operating over a bounded rolling window (“micro-baseline”), producing a reflex deviation score and a constrained routing contract.

Reflex Attention is not a memory tier and does not modify priors. It does not update baselines, commit to consolidation, or trigger high-impact actions on its own. Its sole function is to modulate near-term attention and response shaping—e.g., temporary weighting shifts, clarification prompts, conservative smoothing clamps, or optional handoff to higher governance layers. By decoupling immediate salience from structural update, the reflex layer reduces both overreaction (premature persistence) and underreaction (failure to surface anomalies) under noisy or adversarial conditions. The result is a disciplined architectural primitive that complements constraint-grounded inference, stratified persistence, field dynamics, and constraint deviation governance.

1. Introduction

Stratified agent architectures are designed to remain non-omniscient while preserving time-extended coherence. Constraint-grounded inference restricts what an agent may infer to what its constraint surfaces justify. Stratified memory distinguishes short-term volatility from longer-horizon persistence through gated assimilation and regulated write resistance. Field dynamics models show how distributed instability emerges when amplification dominates dissipation,

often preceded by reference shifts. Constraint deviation governance detects deviation relative to stratified baselines and routes events into governed actions, audit artifacts, and guarded updates.

These layers provide structural discipline, but they leave a practical gap: not all anomalies should be treated as deviation events, and not all salient perturbations should be permitted to influence memory or baseline updates. In real-time operation, systems need a mechanism that can briefly elevate attentional salience under uncertainty—without committing to a structural interpretation. The absence of such a mechanism forces an unhealthy choice: either ignore transient anomalies until they accumulate into structural drift, or treat them as governance events prematurely, increasing false positives and risking uncontrolled assimilation.

This paper specifies a Reflex Attention layer to close that gap. Reflex Attention operates on a bounded rolling window (a micro-baseline) and produces a reflex deviation score intended only to modulate near-term attention and response shaping. Reflex outputs may influence low-impact behaviors—temporary weighting changes, conservative smoothing clamps, clarification prompts, or escalation to review bands—but they are explicitly prohibited from updating baselines, modifying priors, or writing persistent memory. When higher-level governance is required, reflex outputs may be forwarded as an optional input signal to deviation governance layers, where they are evaluated under standard persistence, hysteresis, and audit constraints.

The contribution of this paper is architectural: it formalizes a short-horizon salience primitive that decouples immediate anomaly detection from persistence and governance. This clarifies the division of labor in the Spanda stack: Reflex Attention is pre-structural and non-persistent; deviation governance is structural and routed; memory is tiered and gated. The separation reduces overreach in any single layer and makes system behavior more falsifiable under noisy and adversarial inputs.

2. Terminology and Operational Definitions

All terms are used in an operational and architectural sense.

Reflex Attention

A short-horizon anomaly flag that elevates attentional salience under a bounded rolling window without implying persistence, baseline update, or governance escalation.

Micro-baseline

A bounded rolling reference computed over a short horizon, used solely for reflex scoring. Micro-baselines are not treated as long-horizon baselines and are not eligible for guarded update rules.

Reflex deviation score

A bounded scalar or vector describing the magnitude of short-horizon anomaly relative to the micro-baseline. Reflex deviation is used only for attentional modulation and conservative shaping.

Reflex event

A discrete reflex output produced when reflex deviation exceeds a declared threshold under a persistence rule (if used). Reflex events are non-persistent and time-out by design.

Criticality class

A context label $C \in \{0,1,2,3\}$ that optionally scales reflex priority. Criticality class is declared by the receiving system (e.g., by channel, capability, or protected resource) and does not grant authority to update memory, baselines, or policies.

Criticality multiplier

A manifest-defined multiplier $\kappa(C)$ applied to a reflex score for prioritization. Canonically, $C0$ is informational with $\kappa(C0) = 1.0$.

Reference shift marker

A structural marker indicating a change in interactional orientation or alignment. Reference shift markers are orthogonal to reflex: they may co-occur with reflex spikes but are not implied by reflex salience alone.

Non-persistence guarantee

An architectural constraint that reflex outputs do not commit to durable memory, do not update baselines, and do not trigger irreversible actions without downstream governance.

3. Reflex Layer Contract

Reflex Attention is defined by what it may influence and what it is forbidden to do. It is an attentional modulation layer, not a governance system and not a memory tier.

The reflex layer operates only on constraint-accessible inputs. It may compute short-horizon summaries, rolling window statistics, and bounded feature deltas sufficient to detect anomaly relative to the micro-baseline. The reflex output must be attributable and replayable at the interface level (inputs, parameters, and thresholds declared in the manifest or configuration).

The reflex layer is prohibited from:

- updating long-horizon baselines
- writing to persistent memory tiers
- modifying priors, goals, or policies
- triggering irreversible actions or capability unlocks

The reflex layer may influence:

- attentional weighting of near-term processing (temporary salience modulation)
- conservative smoothing clamps or rate limits on response parameters
- clarification prompts or constraint-seeking queries

- escalation to a review band or forwarding as a signal into downstream deviation governance (optional)

Reflex and structural deviation are separated by design. Reflex can raise salience immediately; structural deviation governance (e.g., CDE) determines whether a turning point has occurred under persistence, hysteresis, and audit constraints.

4. Reflex Scoring and Event Formation

Reflex Attention maintains a micro-baseline for each scoped stream for which reflex is enabled (e.g., per-channel, per-agent, per-scene, or per-capability). The micro-baseline is bounded and short-horizon; it exists to enable immediate anomaly detection and is not a substitute for stratified baselines used by deviation governance.

Micro-baseline construction is non-normative. Canonical implementations include fixed-window estimators (e.g., rolling mean/variance), exponential moving averages with declared decay, and robust short-horizon filters (e.g., median-style estimators) where outlier resistance is desired. The selected estimator and its horizon parameters are declared per stream; micro-baselines remain reflex-only and must not be reused as stratified baselines for governance.

Given constraint-accessible inputs at time t , the reflex layer computes a reflex deviation score R_t relative to the micro-baseline. The score may be scalar or vector-valued, but it must be bounded and interpretable under a declared contract.

Reflex outputs may carry an optional confidence value $c_t \in [0,1]$ indicating reliability of the reflex deviation score under the declared estimator and current input coverage. Confidence is used only for reflex routing decisions (e.g., clamp vs prompt vs forward) and does not by itself justify persistence, baseline update, or governance escalation. Reflex prioritization may incorporate criticality. If a criticality class $C \in \{0,1,2,3\}$ is declared for the active stream or context, the reflex priority may be computed as:

$$P_t = \kappa(C) R_t \text{ where } \kappa(C) \text{ is manifest-defined and } \kappa(C0) = 1.0 \text{ is canonical.}$$

Event formation is optional but permitted. If the architecture benefits from reducing jitter, a reflex event may be emitted only when R_t (or P_t) exceeds a declared threshold for a minimal persistence requirement. Persistence at the reflex layer is bounded and short; it exists only to reduce oscillation and does not create a structural claim. Any reflex persistence parameters (e.g., minimal dwell or timeout) are declared and are not treated as structural evidence absent independent corroboration under downstream governance criteria. Reflex events are time-limited by default and must expire unless re-triggered under the same short-horizon criteria.

Reflex and reference shift markers are orthogonal signals. Reflex encodes immediate salience over a micro-baseline; reference shift encodes structural orientation change. Either may occur with or without the other. When both co-occur, reflex elevates immediate attention while reference shift markers may be forwarded to downstream governance for evaluation under structural criteria.

5. Routing and Downstream Interfaces

Reflex outputs are designed to be low-impact and optionally upstream of governance. Routing is deterministic at the interface level: given the reflex score, confidence (if used), and criticality class, the system selects from a bounded set of admissible responses.

Canonical reflex routing outputs include:

- attentional reweighting (temporary increase in salience for near-term processing)
- conservative clamps (rate limiting, smoothing, or response narrowing under uncertainty)
- clarification prompts (constraint-seeking queries that reduce ambiguity)
- escalation to a review band (human-in-the-loop or supervisory arbitration)
- optional forwarding to deviation governance (e.g., as an auxiliary signal into CDE)

Reflex forwarding does not bypass governance. A downstream governance layer treats reflex as one input stream among others and applies its own persistence, hysteresis, and audit requirements before producing deviation events or governed actions.

Reflex is compatible with stratified memory. It may influence what is attended to, but it does not by itself justify persistence. Memory write decisions remain governed by tier policies and commit criteria defined elsewhere in the stack.

6. Relationship to the Spanda Stack

Reflex Attention aligns with constraint-grounded inference by operating only on constraint-accessible inputs and by avoiding ungrounded structural claims. It does not infer hidden causes; it flags short-horizon anomaly as salience.

Reflex Attention aligns with memory stratification by providing a mechanism that can increase attention without committing to persistence. This reduces pressure to write volatile anomalies into longer-horizon tiers and supports tier discipline under noisy conditions.

Reflex Attention aligns with field dynamics by providing a local salience mechanism that can react before instability propagates. It does not claim to predict cascades; it provides an early attentional flag that can trigger conservative coupling or clarification behavior prior to escalation.

Reflex Attention aligns with deviation governance by clarifying division of labor. CDE detects structural deviation relative to stratified baselines and routes governed events; reflex detects short-horizon anomaly relative to micro-baselines and routes only low-impact attention shaping or optional forwarding signals.

7. Falsifiability and Limitations

Reflex Attention is falsifiable as an interface primitive. The following tests are sufficient to challenge compliance with the specification.

For replayability, a reflex implementation logs at minimum: timestamp, scoped stream identifier, micro-baseline estimator type and horizon parameters (window/decay), a sufficient micro-baseline state summary for recomputation, computed R_p , criticality class C and $\kappa(C)$ if used, threshold(s) and any minimal persistence/timeout parameters if enabled, and the selected routing output (including whether forwarding occurred).

First, non-persistence guarantees must hold: reflex outputs must not write persistent memory, update long-horizon baselines, or trigger irreversible actions without downstream governance. Implementations that allow reflex to directly modify persistence state are non-compliant.

Second, boundedness and expiry must hold: reflex events must be bounded in magnitude under the declared contract and must expire under time-out rules unless re-triggered. Reflex that accumulates without expiry collapses into an uncontrolled memory substitute.

Third, criticality scaling must be deterministic and declared: if criticality is used, $\kappa(C)$ values and routing effects must be manifest-defined. Undeclared criticality behavior undermines auditability.

Fourth, orthogonality to structural deviation must be preserved: reflex salience must not be treated as a turning point claim without downstream governance. When reflex is forwarded, downstream governance must apply persistence, hysteresis, and audit constraints independently.

Reflex Attention does not claim optimal anomaly detection, threat detection, or cascade prediction. It provides a disciplined salience primitive for short-horizon operation and leaves structural interpretation to downstream layers.

8. Conclusion

Reflex Attention fills a layer gap in stratified agent architectures by decoupling immediate salience from persistence and governance. It defines a short-horizon anomaly flag operating over a bounded micro-baseline, with an explicit non-persistence guarantee and deterministic low-

impact routing options. Reflex outputs may influence attention, conservative shaping, and clarification behavior, and may optionally be forwarded into deviation governance, but they do not update baselines, write memory tiers, or trigger irreversible actions on their own.

By formalizing reflex as a pre-structural primitive, the stack becomes more disciplined: constraint-grounded inference bounds what may be inferred, memory stratification governs persistence, field dynamics models collective instability, deviation governance routes structural turning points, and reflex provides a short-horizon attentional trigger that reduces both overreaction and underreaction under noisy or adversarial inputs.