

Inference Residency Creep

DEFINITION

The gradual, unbudgeted expansion of the steady-state inference footprint as models, contexts, and integrations accumulate.

FRAMEWORK FLOW

01 The Condition Every new inference workload inherits operational overhead that never fully exits the environment — canary endpoints, shadow traffic, dependency retention across model generations.	02 The Boundary Whether residency growth is tracked against an explicit ceiling, or each addition is approved individually with nobody tracking the aggregate trend.	03 Failure State Unaudited accumulation — each addition individually justified, the aggregate growing without any team having the authority to address it.	04 Consequence The teams moving fastest — shipping the most valuable work — produce the most creep, because the incentive structure rewards adding, not auditing.
--	---	---	--

Failure State: Unaudited Accumulation



THE DIAGNOSTIC SIGNAL

ADDITION JUSTIFICATION Ask any team whether their specific addition — a canary, a fallback model, a shadow endpoint — was justified. Yes, correctly, every time.	COMPOSITE OWNERSHIP Ask who tracks the combined residency trend across every addition, over time — nobody, because each was approved in isolation.
TRIGGER CONDITION Models are added faster than they're deprecated, and the incentive structure never includes an audit checkpoint.	REMEDIATION SHAPE Endpoint retirement governance, canary retention policies, and dependency lifecycle tracking — a standing check, not a one-time cleanup.

THE THREE WAYS CREEP COMPOUNDS

01 Endpoint Proliferation Each new model deployment creates a new permanent serving floor even if the previous model's was never retired.	02 Observability Accumulation Monitoring and logging infrastructure scales with endpoint count, not traffic — a shadow cost growing with the portfolio.	03 Dependency Retention Token caches and routing layers built for one model generation often persist as dependencies for the next, even after deprecation.
--	--	---

ARCHITECTURAL RELATIONSHIPS

#2 Persistent Inference Residency Stack RELATED MODERATE Residency creep is a growth pattern that emerges inside an unmanaged persistent inference residency stack.	ANCHOR /inference-steady-state-cost/ FRAMEWORK HOME Defines the accumulation pattern that expands inference residency beyond its original operational intent.
--	---

"Every addition to the model portfolio can be individually justified. Nobody can explain the combined residency trend — not because anyone neglected it, but because it was never anyone's to track."