ANALYTIC BYTES

90-Day AI Evaluation Cadence

A practical handoff from AI evaluation diagnosis to operating practice

Chaitanya Ramineni, PhD · Q3 2026 (v3, after external critique)

The cadence is the product.

This is the implementation wrapper for the AI Evaluation Kit. It is intentionally narrow. Pick one pillar, build one instrument, define one threshold, and install one review rhythm that keeps running after the initial push ends.

The operating premise

- **Pick one pillar.** Not three. The kit is finished when the cadence is running. Not when the document has been read.
- **Start with the evidence you can produce.** Plans do not count. Working artifacts do.
- **A threshold is a leadership decision.** The number matters, but the discipline is defending it before it becomes convenient.
- **A dashboard nobody reviews is not an operating system.**

When to use this cadence

Use the 90-day cadence after a leadership team has scored the AI Evaluation Diagnostic and named one pillar to advance. The goal is not Stage 3 maturity. The goal is to move one pillar from informal to documented, or from documented to operating.

Common entry moments where this cadence is the right next move:

- After a vendor pilot decision is made and the org needs to install a review discipline before scaling.
- After a board, funder, or accreditor asks for evidence of responsible-AI practice and a generic principles statement is no longer enough.
- After an internal incident — model output that drifted, subgroup performance that surprised, a teacher / clinician / advisor who pushed back — that exposed missing instrumentation.
- After leadership reads WITSAM or the Validity Ladder and asks *what do we actually do about this on Monday morning?*

The six-phase arc

Timing	Move	What happens
Days 1–14	**Instrument**	Pick the diagnostic question that defines the gap. Build the smallest working instrument that can answer it using current systems and data.
Days 15–30	**Observe**	Run the instrument on real cases. Do not overcorrect too early. Let the pattern show itself across at least two cycles.
Days 31–45	**Define the threshold**	Decide what counts as acceptable, what triggers escalation, and why. Write down the rationale. Let credible colleagues challenge it.
Days 46–60	**Operate**	Put the threshold into a live review rhythm. Assign the owner. Decide what happens when the threshold is breached.
Days 61–75	**Audit**	Sample flagged and unflagged cases. Ask what the instrument caught, what it missed, and where the rule needs revision.
Days 76–90	**Close the arc**	Document the standing review, hand it to the owner, schedule the next review, and make sure the work can continue without the launch team.

What gets produced

Artifact	Standard
Instrument	A dashboard, script, report, regression suite, audit log, or review queue that answers the pillar question on real data.
Threshold	A written decision rule: acceptable range, escalation trigger, rationale, owner, and review rhythm.
Standing review	A scheduled meeting or workflow where the instrument is reviewed and decisions are made.
Owner	A named person with enough authority to keep the review alive and escalate when needed.
Handoff note	A short operating memo explaining the metric, threshold, review rhythm, and next maturity move.

Examples by pillar

Behavior example

Gap: The organization uses an LLM judge but has never audited it. The 90-day move is a judge reliability audit: five to ten representative cases, blind human review, LLM judge scoring, bias pattern analysis, and a decision on whether to keep, wrap, or replace the judge.

Signal Quality example

Gap: The evaluation set is treated as valid but has never been compared to production traffic. The 90-day move is an eval-set realism check: profile production cases, compare against the evaluation set, identify underrepresented cases, and revise headline claims accordingly. *Construct-drift sub-check: if the construct (depression severity for a defined population, student engagement for a specific cohort, fairness for a defined subgroup) has had a definitional update in the literature, in clinical guidance, in state standards, or in your operating context in the last twelve months, the realism check has to absorb that shift.*

Governance example

Gap: Humans can override the model, but nobody reviews override patterns. The 90-day move is a discrepancy policy: define what counts as disagreement, instrument overrides, review patterns by subgroup/use case, and name an escalation owner.

Vendor-evaluation example

Gap: An AI vendor pitch is in front of leadership. The vendor carries a responsible-AI badge or framework alignment (Polaris Principles, EdSafe AI Alliance, Common Sense AI rating, ORCHA review). The 90-day move is to install a vendor-claim stress-test: walk the twelve diagnostic questions against the vendor's published claims; identify which can be verified from public evidence vs. require vendor disclosure; produce a procurement decision memo with the answers documented and the open questions named. The review rhythm after deployment becomes the standing operating review for whether the vendor's claims continue to hold. *For procurement-grade evidence request sheets, red flags, and contract language: see the AB Vendor Evidence Pack (companion download).*

What this is not

- It is not a full responsible-AI program. Frameworks like NIST AI RMF, NAM AI Code of Conduct (May 2025), APA Ethical Guidance, WHO recommendations, EdSafe AI Alliance, Common Sense AI ratings, and ORCHA reviews already operate at the framework / seal / certification layer. This cadence operates beneath them.
- It is not a benchmark-selection exercise.
- It is not a replacement for technical evaluation tooling (HELM, MMLU, evaluator LLMs, internal regression suites).
- It is not a six-month strategy deck.
- It is the operating bridge between evidence and decisions.

Advisory shape

Phase	Analytic Bytes role	Client role
Diagnostic	Facilitate scoring, surface gaps, name the highest-leverage pillar.	Bring product, data, domain, and risk perspectives to the table.
Instrument	Design the minimum viable measurement artifact and threshold logic.	Provide access to workflows, sample cases, and existing evidence.
Operate	Install the review rhythm and decision rules.	Run the rhythm with named owners.
Handoff	Document the operating model and next maturity move.	Own the rhythm after the first arc closes.

The closing standard

A 90-day arc is successful when a leader can answer four questions without hunting through a deck:

1. What are we measuring?
2. What threshold matters?
3. Who reviews it?
4. What decision follows when the evidence changes?

That is what turns AI evaluation from a technical activity into a decision system.

Companion reads — Analytic Bytes Library

- **"What is this system actually measuring?" (WITSAM)** — conceptual anchor.
- **The Validity Ladder** *(artifact)* — five rungs of evidence the cadence walks.
- **Reliability vs Validity** *(artifact)* — Q1 and Q5 made visual.
- **Fair for Whom?** *(artifact)* — Q9 made visual.
- **"It's not a communication issue. It's a blown assignment." (Cross-functional is a football play)** *(field note)* — the alignment / assignment / execute spine. The cadence is the play running; the diagnostic is whether each phase actually lands.
- **The Validity Layer Beneath Responsible AI** *(essay, forthcoming)* — where this cadence sits in the framework landscape.
- **The construct keeps moving** *(essay, forthcoming)* — why the construct-drift sub-check matters.
- **AB Vendor Evidence Pack** *(companion download, forthcoming)* — procurement-grade evidence request sheets, red flags, contract language, local pilot validation plans.

Contact: chaitanya.ramineni@gmail.com · analyticbytes.systems

Analytic Bytes · From Fragmented to Decision-Ready · Q3 2026 v3