ANALYTIC BYTES

The AI Evaluation Kit

A buyer-side validity stress test for AI vendor claims

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

AI evaluation is not just a model-benchmarking problem. It is a measurement, governance, and operating-cadence problem.

Most organizations can now generate AI outputs. Fewer can say, with evidence, whether those outputs are good enough to support real decisions under real operating pressure. This kit is for that gap.

Why this exists

The AI evaluation conversation is moving fast, but a lot of it still lives too close to the model: accuracy, latency, token cost, hallucination rate, leaderboard performance. Those matter. They are not enough.

The harder question is whether the system is measuring what the organization thinks it is measuring, for the people it serves, with decision rules that hold up when the system is stressed.

> **The working thesis**
>
> AI evaluation under operational conditions is a measurement problem, not just an engineering problem. The tooling has changed. The discipline has not.

Where this fits

Frameworks, seals, and ratings can tell you whether a vendor speaks the responsible-AI language. They do not tell you whether *this* AI claim is valid for *your* decision, *your* population, and *your* workflow.

The 2026 floor — NIST AI RMF, NAM AI Code of Conduct (May 2025), APA Ethical Guidance (2025-2026), WHO Jan 2026 workshop, EU AI Act on the regulatory side; Polaris Principles, EdSafe AI Alliance, Common Sense Education AI ratings, ORCHA on the vendor-seal side; Lumara's five-question rubric on the buyer-rubric side — names what responsible AI should include. This kit names what evidence a buyer should demand, what threshold should be set, who reviews it, and when to stop or scale.

Use the framework as the floor. Use this kit to decide what to require before signing, scaling, or renewing.

When the vendor says they're responsible-AI-compliant

The kit's questions translate directly into vendor stress-tests for three concrete buyer moments.

Higher-Ed CIO evaluating an AI-augmented student success platform

(the EAB Navigate AI / Civitas Learning AI class). The pitch says *"identifies at-risk students early so advisors can intervene, aligned with [framework X]."*

The blown assignment to prevent: the platform flags amber on a first-generation student, the advisor's caseload makes a same-day reach impossible, and the student withdraws while the system marks them as "outreach pending."

Questions to ask the vendor before signing: - *Q5 sharper version:* "What does 'at-risk' mean here — at risk of not enrolling next term, not graduating in six years, not passing a specific course? Each is a different construct." - *Q7 sharper version:* "What external outcome should move if this AI is working? Show the evidence — particularly: does the risk prediction correlate with actual outcomes when no intervention happens, or only when intervention does happen?" - *Q9 sharper version:* "Which subgroup-by-use-case cells (first-gen × STEM major; Pell-eligible × first-term transfer; international × upper-division) have enough volume to trust the estimate, which do not, and what decision rule applies when evidence is thin?" - *Q11 sharper version:* "What second-order effects appear when advisors act on the flags — over-monitoring, student gaming, advisor burnout, dropout from over-intervention?"

K-12 Superintendent piloting a teacher-productivity AI

(the MagicSchool / Khanmigo class). The pitch says *"saves teachers time, aligned with EdSafe AI Alliance principles."*

The blown assignment to prevent: the AI generates a lesson plan that hallucinates a historical fact or misaligns with state curriculum, the teacher pastes it directly into Monday's class, a parent flags it, and the district doesn't have a policy for who's accountable.

Questions to ask the vendor before signing: - *Q5 sharper version:* "What exactly is the construct — time saved, lesson quality, standards alignment, differentiation quality, teacher workload, or student learning? Which were measured, in what district context, and what got worse while time improved?" - *Q4 (judge-of-judge):* "If the AI is checking other AI-generated content, has that judge been audited for bias on K-12 instructional tasks?" - *Q3 sharper version:* "What misuse cases have you tested specifically for K-12 instructional workflows — prompt leakage, standards drift, age-appropriateness drift — and which ones still fail?" - *Q12:* "What would trigger the district to revisit this — quarterly review, a state-standards revision, a teacher-reported drift, a parent complaint?"

Behavioral Health Clinical Director considering an AI screening tool

(the Wysa / Woebot class for screening or AI triage tools more broadly). The pitch references Polaris Principles or NAM Code of Conduct.

The blown assignment to prevent: the AI flags suicide-risk on a 17-year-old in an under-resourced clinic, the workflow doesn't define a same-day warm-handoff owner, the flag sits in a queue, and the patient is missed.

Questions to ask the vendor before signing: - *Q5 sharper version:* "Which construct does the AI claim to measure — depression severity, suicide risk, treatment readiness? Each has different validity standards in the published literature, different escalation duties, and different false-negative risks." - *Q9 sharper version:* "Which patient-by-symptom-class subgroups have enough volume to trust the model's performance estimate? Racial and ethnic minorities, non-English speakers, LGBTQ+ youth specifically — what was measured, on what sample, and what's the decision rule when subgroup evidence is thin?" - *Q10:* "When the model and a clinician disagree, what happens, who decides, and is that rule audited?" - *Q11 sharper version:* "What clinical action follows the score, and has *that action* been validated? A depression-severity score, a suicide-risk triage flag, and a treatment-readiness score create different duties and different liability exposure."

The three pillars

Pillar	Question it answers	What failure looks like
Model Behavior	Does the model do what you say it does — consistently, and under pressure?	The model looks good in demos but fails under repeated sampling, edge cases, model upgrades, or adversarial use.
Signal Quality	Is the model measuring what you say it is measuring — and does that match the real world?	The metric is clean, but the construct is weak. The benchmark does not match production. The score does not move the outcome.
Governance	Who does this work for, who decides, and what happens when it goes wrong?	Aggregate metrics hide subgroup harm. Overrides are informal. Consequences accumulate without review. Nobody knows when to roll back.

The twelve diagnostic questions

Pillar 1 — Model Behavior

- **Q1. Reference agreement under stochasticity:** What is agreement with human, expert, or reference judgments across repeated samples — not just one lucky run?
- **Q2. Release-tracked stability:** Show performance before and after the last three model, prompt, retrieval, or policy updates. What changed, and what release gate caught it?
- **Q3. Workflow-specific gaming:** What misuse cases have you tested that are specific to *this workflow* (prompt leakage, answer laundering, rubric chasing, crisis-language manipulation, reassurance loops, age-appropriateness drift), and which ones still fail?
- **Q4. Judge-of-judge reliability:** If an LLM is scoring, ranking, or grading, has that judge been audited for bias on this specific task?

Pillar 2 — Signal Quality

- **Q5. Construct definition and drift:** What construct does the AI claim to measure, who defines it, and what would make that definition obsolete? Review triggers include changes in clinical guidance, state standards, institutional policy, population served, workflow use, or fairness expectations. *A validity claim made at deployment does not survive material changes in the construct, population, workflow, or policy environment automatically.*
- **Q6. Eval-set realism and contamination:** Was the evaluation set built from cases like yours, after deduplication, and separated from training, fine-tuning, and prompt-development examples?
- **Q7. External alignment:** What external outcome, criterion measure, or human expert judgment should move if this AI is working — and what evidence shows that it does? *Distinguish: does the score correlate with outcomes when no intervention happens, or only with outcomes after intervention does happen?*
- **Q8. Construct undercoverage:** Which important parts of the construct are *not* captured, and where could that omission create a bad decision?

Pillar 3 — Governance

- **Q9. Subgroup-by-use-case performance:** Which subgroup-by-use-case cells have enough volume to trust the model's estimate, which do not, and what decision rule applies when evidence is thin?
- **Q10. Discrepancy policy:** When the model and a human disagree, what happens, who decides, and is the rule audited?
- **Q11. Consequence audit:** What second-order effects are appearing because the model is now part of the workflow — over-monitoring, gaming, frustration, alert fatigue, deskilling, workload distortion?
- **Q12. Sunset criteria:** What conditions would trigger review, retraining, rollback, or retirement?

What the diagnostic does not cover — additional vendor-pitch questions

The twelve questions are a buyer-side validity stress test. They do not cover the full procurement surface. For the next layer — vendor evidence requests, contract language, incident response, and local pilot validation — see the *AB Vendor Evidence Pack* (companion download, forthcoming).

A buyer evaluating a 2026 AI vendor will also want to ask:

- **Data use and model training rights.** Can the vendor use client prompts, student work, therapy-adjacent text, or educator content for training, evaluation, product improvement, or human review? Under what NDA?
- **Evidence access and auditability.** Will the vendor share test plans, subgroup tables, incident history, release notes, known limitations, and eval methodology under NDA?
- **Human factors and adoption burden.** What training is required, who gets deskilled, who gets more work, and what failure mode appears under workload pressure?
- **Incident response.** What counts as an AI incident, how fast must the vendor notify the buyer, and what evidence must be preserved?
- **Contractual commitments.** Which claims become contractual obligations? If the vendor says "monitored for bias," does the contract require subgroup reporting, cadence, and remediation?
- **Local validation before scale.** What must be true *in this institution, district, or agency* before expansion?

These are procurement questions. The diagnostic above is the validity stress test that has to be passed before procurement questions are worth asking.

The maturity read

The point is not to rate everything optimistically. The point is to find the weakest link before the system finds it for you.

Stage	Behavior	Signal Quality	Governance
0 — Vibes	Demos, impressions, no reference standard.	Implicit: no written construct definition.	Trust-the-engineer: no subgroup metrics or rollback plan.
1 — Eyeballed	Spot checks, anecdotal failure modes.	Defined: construct written, eval set documented.	Documented: policies exist, but not yet operating.

Stage	Behavior	Signal Quality	Governance
2 — Measured	Reference set, threshold, drift cadence.	Triangulated: production comparison, external outcome, contamination audit.	Operating: subgroup review, discrepancy policy, consequence audit, sunset criteria.
3 — Argued / Adversarial	Red-teaming, failure modes used in release gates.	Argued: validity argument, effect-size evidence, limitations disclosed.	Governed: exec visibility, external review, criteria triggered and acted on.

How to use it

- **Read once:** Use the executive kit to align the leadership team on what AI evaluation has to cover.
- **Score honestly:** Bring the one-page diagnostic to a leadership meeting. Score each pillar based on evidence, not intent.
- **Pick one pillar:** Do not try to mature everything at once. Name the highest-leverage pillar to advance in 90 days.
- **Run the cadence:** Instrument, observe, define the threshold, operate, audit, and hand off to an owner.

What Analytic Bytes brings

Analytic Bytes helps learning and health organizations turn AI pilots into evidence-backed operating decisions: what to measure, what threshold matters, who reviews it, and when to stop. The work is not a strategy deck. It is a decision system: evidence, thresholds, owners, and a standing review rhythm.

Companion reads — Analytic Bytes Library

The kit operationalizes a longer argument developed across the AB Library.

- **“What is this system actually measuring?” (WITSAM)** — the conceptual anchor. The kit operationalizes the essay.
- **The Validity Ladder** *(artifact)* — five rungs of evidence. The kit’s Pillar 2 sits on this ladder.
- **Reliability vs Validity** *(artifact)* — the four-target view. Q1 and Q5 are this contrast made operational.

- **Fair for Whom?** *(artifact)* — fairness as validity asked one subgroup at a time. Q9 is this artifact in question form.
- **Measurement = Diagnostics** *(artifact)* — vocabulary bridge for cross-sector buyers.
- **"It's not a communication issue. It's a blown assignment." (Cross-functional is a football play)** *(field note)* — the brand-spine field note on alignment, assignment, and execution. The vendor-pitch vignettes in this kit are the validity equivalent: the questions that prevent the blown assignment when the AI signal arrives at the desk.
- **The Validity Layer Beneath Responsible AI** *(essay, forthcoming)* — long-form articulation of where AB sits in the framework landscape.
- **The construct keeps moving** *(essay, forthcoming)* — why validity is a governance pillar, especially in mental health.
- **AB Vendor Evidence Pack** *(companion download, forthcoming)* — evidence request sheets, go/pilot/no-go memo templates, red-flag lists, local pilot validation plans, contract language prompts. The procurement-grade companion to this kit.
