

ANALYTIC BYTES DIAGNOSTIC

AI Evaluation Diagnostic

Three pillars. Twelve questions. One leadership meeting. One 90-day operating move.

Pillar	Diagnostic question	Stage 0 — Vibes	Stage 1 — Eyeballed	Operating standard (Stage 2 / 3)
Behavior	1. Agreement under repeated samples	No reference set	Spot checks	Reference set + threshold + repeated samples
Behavior	2. Release-tracked stability	No drift check	Upgrade checks only	Release-gated: performance before/after last 3 updates documented
Behavior	3. Workflow-specific gaming	Nobody has tried to break it	Anecdotal failures	Workflow-specific misuse cases tested + documented failure modes
Behavior	4. Judge-of-judge reliability	LLM judge trusted by default	Manual spot check	Bias audit on this task + cross-validation
Signal	5. Construct definition and drift	No shared definition	Written definition	Metric-construct map defended + review triggered by changes in clinical guidance, state standards, policy, population, workflow, or fairness expectations

Pillar	Diagnostic question	Stage 0 — Vibes	Stage 1 — Eyeballed	Operating standard (Stage 2 / 3)
Signal	6. Eval realism + contamination	Eval set assumed valid	Eval set documented	Built from cases like yours, deduplicated, separated from training/fine-tuning/prompt-development External criterion or outcome moves with the score; intervention-confounding addressed Undercovered construct parts named + decision-risk mapped Subgroup-by-use-case cells with adequate n identified; decision rule named when evidence is thin Written escalation policy + audit Quarterly review for over-monitoring, gaming, alert fatigue, deskilling, workflow distortion
Signal	7. External alignment	Internal metric only	Early correlation	
Signal	8. Construct undercoverage	Limitations unstated	Known internally	
Governance	9. Subgroup-by-use-case performance	Aggregate metrics only	One-off subgroup analysis	
Governance	10. Discrepancy policy	Ad hoc overrides	Informal rule	
Governance	11. Consequence audit	No washback review	Anecdotes only	

Pillar	Diagnostic question	Stage 0 — Vibes	Stage 1 — Eyeballed	Operating standard (Stage 2 / 3)
Governance	12. Sunset criteria	No rollback conditions	Theoretical answer	Monitored criteria + named owner

How to score honestly

- Evidence, not intent: plans do not count.
- Use three scorers: product/program, data/engineering, research/evaluation/risk.
- Lowest credible cell in a pillar is the pillar score.
- Output: one score per pillar + one named pillar to advance in 90 days.

When a vendor pitches you

These twelve questions are also the validity questions to ask when a vendor claims responsible-AI-compliance. The vendor's framework alignment (Polaris Principles, Ed-Safe AI Alliance, Common Sense Education AI ratings, ORCHA review, NAM / APA / WHO references) is the floor — what *responsible AI should include*. These twelve are what to ask *beneath* the floor — what evidence makes the claim valid for *your* decision, *your* population, *your* workflow.

For the next layer — vendor evidence request sheets, red flags, contract language, incident response, local pilot validation plans — see the *AB Vendor Evidence Pack* (companion download, forthcoming).

Leadership meeting output

Behavior score	_____
Signal score	_____
Governance score	_____
Named pillar to advance	_____
First 90-day move	Instrument → Observe → Define threshold → Operate → Audit → Handoff
Owner	_____
Next review	_____

Companion: WITSAM, the Validity Ladder, and “It’s not a communication issue. It’s a blown assignment.” — at analyticbytes.systems/library.

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