CACTUS

EVALUATION AS A SERVICE

Validation Readiness Assessment

A no-cost diagnostic of how ready your AI evaluation programme is to withstand scrutiny from regulators, customers and your own board.

ILLUSTRATIVE SAMPLE Not a live client report. Composite example, fictional organisation.

PREPARED FOR

Meridian Diagnostics AI

Clinical decision-support models | Series B | Boston, MA

ASSESSMENT WINDOW	EVALUATION SURFACE	READINESS BAND	REPORT DATE
12 working days	3 production models	Developing	March 2026

WHAT THIS REPORT GIVES YOU

- A readiness score across the six dimensions that determine whether an evaluation programme holds up under audit.
- The specific gaps standing between your current evaluation practice and defensible, evidence-backed assurance.
- A recommended next step, scoped and costed, with no obligation to take it with us.

How to read this report

The Readiness Assessment is a diagnostic, not an audit. It tells you where your evaluation programme is strong, where it is thin, and what it would take to make it defensible. It is deliberately short. Everything in it should be arguable in a room with your technical and compliance leads.

WHAT WE ASSESSED AGAINST

Six dimensions drawn from two decades of evaluation work across regulated publishing, pharma and enterprise AI deployment. Each is scored 0 to 5 against observable evidence, not stated intent. **A claim without an artefact behind it scores as absent.**

INPUTS TO THIS ASSESSMENT

Structured intake	18-question readiness questionnaire completed by the VP Engineering and Head of Regulatory.
Working sessions	Two 60-minute sessions with the ML platform team and the clinical safety lead.
Artefact review	Model cards for three production models, current annotation guidelines, two internal validation memos, sample evaluation exports.
Output sampling	400 randomly drawn model outputs re-reviewed by two CACTUS domain experts against your existing rubric.

SCORING BANDS

0 – 1	**Absent**	No process, no artefact, no owner.
2	**Emerging**	Practice exists informally. Not documented, not repeatable.
3	**Developing**	Documented and repeatable. Not yet independently verifiable.
4	**Established**	Verifiable, with evidence trail. Gaps are known and managed.
5	**Defensible**	Withstands external audit without remediation.

Readiness summary

OVERALL READINESS

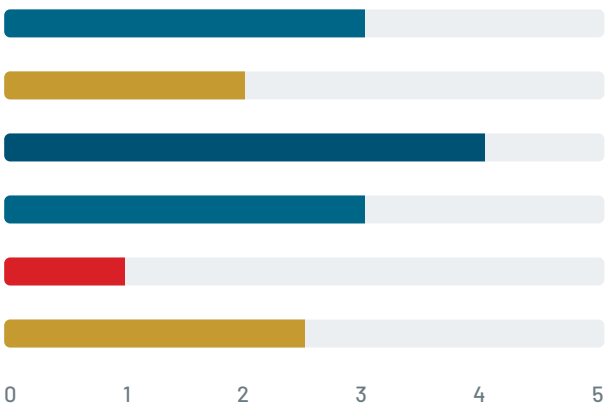
2.6/5

DEVELOPING

HEADLINE FINDING

Meridian's evaluation practice is technically competent and materially undocumented. The team knows its models are performing; it cannot yet show a third party why. **The gap is not capability, it is evidence.** Two of the three production models would not survive a payer or FDA-adjacent evidence request in their current state.

DIMENSION SCORES



THE DIMENSION THAT DECIDES THE REST

Risk, compliance and audit trail scores 1. There is no versioned record linking a model release to the evaluation evidence that cleared it. Every other score in this report is therefore unverifiable to an outside party, including the ones that are strong. Fixing this single dimension raises the defensibility of the whole programme more than any other intervention available to you.

Findings by dimension

Each dimension carries what we observed, the risk it creates, and the specific move that would raise the score.

Evaluation objectives and success criteria 3

OBSERVED	Success thresholds exist for accuracy and latency and are consistently applied. They are set by the ML team alone.
RISK	No clinical or regulatory stakeholder signs off on what 'good enough' means, so thresholds cannot be defended as fit for clinical purpose.
MOVE	**Introduce a documented threshold-setting ritual with named clinical sign-off per model release.**

Ground truth and reference data 2

OBSERVED	Reference sets were assembled ad hoc by two engineers during the 2024 build and have not been refreshed.
RISK	Reference data has drifted from the current patient population. Two of three models are being validated against a distribution they no longer serve.
MOVE	**Rebuild a versioned, clinically reviewed gold-standard set with documented provenance and a refresh cadence.**

Domain expertise coverage 4

OBSERVED	Three board-certified clinicians review edge cases weekly. Genuine depth here and the strongest part of the programme.
RISK	Coverage is concentrated in two specialties and depends on named individuals. No documented escalation path if either leaves.
MOVE	**Formalise the reviewer panel with defined coverage per specialty and a bench of vetted substitutes.**

Findings by dimension continued

Output structure and pipeline integration 3

OBSERVED Evaluation results are exported as structured JSON and land in the data warehouse.

RISK Schema is undocumented and has changed twice without versioning, so historical comparisons are unreliable.

MOVE **Publish a versioned output schema and enforce it at write time.**

Risk, compliance and audit trail 1

OBSERVED Evaluation happens. Evidence of it does not persist in any retrievable, tamper-evident form.

RISK No release-to-evidence linkage. In a regulatory or payer evidence request you would be reconstructing history from Slack threads and notebooks.

MOVE **Stand up an evaluation record of decision per model release: inputs, reviewers, thresholds, outcome, sign-off, immutably stored.**

Operating model and scale readiness 2.5

OBSERVED The current process works at three models and roughly 40 reviewer hours a month.

RISK It is entirely manual and owned by one platform engineer. At the eight models planned for FY27 it breaks.

MOVE **Separate the evaluation function from the build team and move first-pass review to an automated pass with expert validation on exceptions.**

Where we would start

Six dimensions scored low is not six problems to solve at once. The order matters more than the effort. This is the sequence we would follow, and it holds whether the work is done internally, by us, or by someone else.

1
FIRST

Close the evidence gap

Stand up an evaluation record of decision for the next model release. Inputs, reviewers, thresholds applied, outcome, sign-off, stored immutably.

Nothing else in this report can be demonstrated to an outsider until this exists. It is also the smallest piece of work on the list.

4 TO 6 WEEKS · ONE ENGINEER PART-TIME

2
SECOND

Rebuild the reference set

Reconstruct the gold-standard data with clinical review and documented provenance, then set a refresh cadence against your current patient population.

Two of three models are being validated against a distribution that has moved. Every score above this line inherits that error.

6 TO 10 WEEKS · CLINICAL REVIEWER TIME IS THE CONSTRAINT

3
THIRD

Version the output schema

Publish the evaluation output schema, enforce it at write time, and backfill version tags on historical exports where feasible.

Cheap to do and it makes your existing two years of evaluation data comparable again rather than indicative only.

2 TO 3 WEEKS

4
FOURTH

Separate evaluation from build

Move first-pass review off the platform engineer who owns it today. Automate the routine pass, route exceptions to the expert panel.

The current model works at three models and fails at eight. This is the only item on the list that is about FY27 rather than today.

PLAN NOW · EXECUTE BEFORE MODEL FIVE

A NOTE ON SEQUENCE

Steps one and three are inexpensive and can run in parallel. We would not begin step two until step one is in place, or the rebuilt reference set inherits the same traceability problem it is meant to solve.

Method and terms

HOW THE SCORES ARE PRODUCED

Two CACTUS assessors score each dimension independently against a fixed rubric using only observable artefacts. Divergence greater than one point is resolved in a moderation call, and the moderated score is the one reported. Where an organisation describes a practice but cannot produce evidence of it, the dimension is scored as though the practice is absent. **This is deliberately unforgiving and it is the reason the assessment is useful.**

WHAT THE ASSESSMENT DOES NOT DO

- **It does not certify your models.** No readiness score constitutes regulatory clearance of any kind.
- **It does not evaluate model performance.** It evaluates the evidence you hold about model performance.
- **It does not require a commercial commitment.** Roughly a third of assessments we run close with a recommendation the organisation executes internally.

ABOUT THIS SAMPLE

Meridian Diagnostics AI is a fictional organisation. The findings shown are a composite drawn from anonymised patterns across real assessments, assembled to illustrate the format and depth of the deliverable. No client data appears in this document.