AI Ops Readiness Scorecard

Score your AI feature in 5 minutes against the same 6-component framework The Content Matrix uses with paying clients.

> Per 2026 industry analysis, AI agents with **full evaluation coverage** had a **9% production rollback rate** over the prior year. Agents **without** had **47%**. Same models. Same budget. 5× difference in whether the feature stayed up — explained almost entirely by the operational layer underneath.
>
> Source: digitalapplied.com 2026 enterprise data points · Fiddler AI 2026 production analysis

How to score: For each question, check the box that best matches your current state. Sum your points at the end (out of 24). Then map your score to the tier table on page 4.

COMPONENT 1 OF 6

Eval Harness

Scoring AI output against representative inputs on every prompt or model change. The single most predictive component of whether a feature stays in production.

QUESTION 1

Do you have at least 10 representative input cases with asserted expected properties for this AI feature?

☐ No	☐ Partial / informal	☐ Yes, automated
0 PTS	1 PT	2 PTS

QUESTION 2

Do those eval cases run **automatically** on every prompt or model change?

☐ No	☐ Sometimes / manually	☐ Yes, every change
0 PTS	1 PT	2 PTS

COMPONENT 2 OF 6

Verification Layer

Checking each factual claim against authoritative source data. Catches confident hallucinations before they ship.

QUESTION 3

Does every factual claim in your AI's output get checked against authoritative source data — not just the model's own recall?

☐ No	☐ Some claims	☐ Yes, every claim
0 PTS	1 PT	2 PTS

QUESTION 4

Is that verification check **independent** of the model that produced the claim (separate model, retrieval, or rules-based)?

☐ No	☐ Partially	☐ Yes, fully independent
0 PTS	1 PT	2 PTS

COMPONENT 3 OF 6

State + Idempotency

Making sure retries and partial failures don't double-fire side effects. Prevents duplicate emails, double charges, repeated posts.

QUESTION 5

Does every side-effecting action your AI takes (send, post, charge, write) have a unique idempotency key?

☐ No	☐ Some actions	☐ Yes, all actions
0 PTS	1 PT	2 PTS

QUESTION 6

Can a retry of a partially-failed action be triggered without producing duplicate side effects?

☐ No / unsure	☐ Sometimes	☐ Yes, guaranteed
0 PTS	1 PT	2 PTS

COMPONENT 4 OF 6

Cost + Rate Guards

Hard spend caps that fail the workflow closed when hit. Prevents the runaway loop billing four figures overnight.

QUESTION 7

Is there a hard per-run spend cap that fails the workflow **closed** when hit?

☐ No	☐ Soft alert only	☐ Yes, hard close
0 PTS	1 PT	2 PTS

QUESTION 8

Is there a per-day or per-account spend cap with alerts firing **before** the cap is hit?

☐ No	☐ Cap only, no alerts	☐ Yes, both
0 PTS	1 PT	2 PTS

COMPONENT 5 OF 6

Observability

Structured logs of every model call, tool call, and decision — enough to reconstruct any single run during a review.

QUESTION 9

Can you reconstruct every model call, tool call, and decision a single AI run made from your logs?

☐ No	☐ Some runs	☐ Yes, every run
0 PTS	1 PT	2 PTS

QUESTION 10

Are those logs structured (JSON) and queryable — not just stdout dumps?

☐ No	☐ Partially	☐ Yes, fully
0 PTS	1 PT	2 PTS

COMPONENT 6 OF 6

Approval Gate

Explicit human approval before any irreversible action — enforced in code, not just policy.

QUESTION 11

Does every irreversible action (publish, send, charge, delete) require explicit human approval before firing?

☐ No	☐ Some actions	☐ Yes, all actions
0 PTS	1 PT	2 PTS

QUESTION 12

Is the approval gate enforced **in code** — not just in policy — so it can't be bypassed by mistake?

☐ No	☐ Policy only	☐ Yes, in code
0 PTS	1 PT	2 PTS

What Your Score Means

22–24 TIER 1

Production-grade

You're in the 9% group. The 6 components are in place. Next move: automate the build playbook for new features so this stays true as you scale.

17–21 TIER 2

One incident from a rollback

1–2 components are weak. The questions you scored 0 or 1 on are your rollback risk. Most common gaps at this tier: cost guards and observability.

12–16 TIER 3

High-risk

You're in the 47% rollback population. Half your ops layer is missing. Highest-leverage single fix: an eval harness — the most predictive component in the 2026 data.

0–11 TIER 4

Pre-production

Don't ship to production until you minimum-deploy eval harness, idempotency, and observability. The gap to the 9% group is structural — there is no shortcut.

How to use this: The three components where you scored lowest are your priority build list — in that order. Fix the lowest-scoring component first. Re-take the scorecard after each fix to track progress.

Want a 15-minute review with the TCM Founder?

We'll score one of your features together and give you a build order for your stack. No pitch.

calendly.com/thecontentmatrix/blueprint-walkthrough

Sources for the 9% vs 47% claim: digitalapplied.com 2026 enterprise data points (analysis of 120+ AI agent deployments) · Fiddler AI 2026 ("AI Agent Failure Rate" production analysis). Same sources cited in our AI Ops Layer blog at thecontentmatrix.net/blog-ai-ops-layer.

About this scorecard: The 6-component framework The Content Matrix uses with paying clients to score AI features before they ship. Each component maps to a documented failure mode in the 2026 data. Re-score after every change to the feature.
