AI Reliability Scorecard

Baseline your production AI: reliability • risk • ROI in 10 minutes.

What this is

A one-page framework to quickly assess an AI workflow's production readiness. Use it to identify where you're leaking reliability (hallucinations, escalations, compliance exposure) and what to fix first.

Score 1–5 across five dimensions

- Task success rate (pass/fail on real examples)
- Critical error rate (harmful, incorrect, or policy-violating outputs)
- Escalation rate (how often humans must intervene)
- Cost per successful task (tokens + tools + human time)
- Observability & governance (logs, evaluation, audit trail, controls)

Failure-mode checklist

- Hallucinated facts / invented citations
- Tool misuse (wrong API calls, unsafe actions)
- Policy/compliance leakage (PII, regulated claims)
- Non-deterministic behavior across the same input
- Silent degradation after model/KB changes

What good looks like (targets)

- ≥90% task success on a representative eval set
- ≤1–2% critical errors (target: near-zero for regulated flows)
- Escalations trending down month-over-month
- Stable cost per successful task with guardrails
- Documented go-live checklist + monitoring + incident playbook

Next step

Book a 30-minute AI Risk Audit to map one workflow, define pass/fail, identify top failure modes, and outline a 10■day Reliability Sprint plan.