Agent QA Toolkit
Portable Evidence Packs for tool-using AI agents

- Offline, self-hosted, no data egress
- CI gate decisions + security signals
- Target: platform/QA engineers shipping agents in production

Problem

- Incident handoff is still screenshots + partial logs
- Sharing runs outside your tracing UI is slow or impossible
- Tool I/O and retrieval context are scattered
- Secrets/PII make sharing risky

Solution

- One run → one portable Evidence Pack
- report.html (human) + compare-report.json (CI contract)
- assets/ + manifest.json with sha256 integrity
- Strict verify: portability + integrity checks

How It Works

- Runner executes cases against the agent
- Evaluator diffs baseline vs new
- Outputs Evidence Pack (offline, shareable)
- CI consumes per-case gate decisions

What You Get

- Regression diffs per case
- Root cause attribution (RCA)
- Security scanners (PII, injection, exfiltration, risk)
- Self-contained, portable report directory

Why Not Promptfoo / DeepEval / LangSmith
- They focus on model quality and hosted traces
- We focus on agent behavior + offline handoff
- Evidence Pack is attachable to tickets
- No SaaS dependency

Deployment

- Self-hosted only
- Works in air-gapped environments
- No accounts or external storage
- You control retention and sharing

Pricing
- Free: open-source core
- Services-first: onboarding + CI setup + custom rules
- Pro (later): rules library + trending
- Enterprise: advanced redaction + support SLA

Use Cases

- Release gates before deploying a new agent version
- Incident support bundles for customers/vendors
- Evidence packs for compliance teams

Call to Action

- Pilot: 5–10 teams
- We can connect your agent in 1 day
- Start with a demo bundle and CI gate