W5: Agent Evaluation and Operations Plan (capstone)

Evaluating and Operating Agentic AI Systems: Graders, Trajectories, Red-Teaming, and Production Monitoring (FAST-13) · Lab 5: Capstone, the Agent Evaluation and Operations Plan

Name: _____ Date: _____

1. **Agent and purpose.** What it does, for whom, which tools, and what it can change (write access).
2. **Top failure modes.** From error analysis, with counts and severity.
3. **Eval suite.** For each failure mode: grader type, input, pass condition, and validation (for a judge, the kappa against humans).
4. **Component coverage.** Router, tool arguments, skills, trajectory (match mode and why), outcome check.
5. **Reliability and cost.** Trials per task, pass^k target, cost per success budget, p95 latency target.
6. **Security.** The top 3 OWASP agentic risks for this agent, the red-team cases, the guardrails, and their measured or target miss and false-alarm rates.
7. **Tracing and privacy.** What is captured, redaction, retention period, who can access traces, and how this meets 201 CMR 17.00.
8. **Online evaluation and monitoring.** Which graders run on all traffic and which on a sample (with the rate), dashboards, alerts with thresholds, and SLOs.
9. **Release process.** CI gate datasets and thresholds, shadow, canary, and A/B steps, rollback and kill switch, and the named sign-off owner.
10. **Incident loop and documentation.** Postmortem trigger, how failures become test cases, eval report and system card owners, and NIST AI RMF "Measure" mapping.

Completion criteria. All ten sections present; meets the course rubric (failure-driven, grader quality, coverage, reliability and cost, security, operations); presented and peer reviewed. One resubmission is allowed within the course window.

Take-home. One-page agent evaluation and operations checklist (handout), which summarizes the ten sections above.