Take-Home Checklist: Agent Evaluation and Operations

FAST-13 Evaluating and Operating Agentic AI Systems · 7formations · Instructor: Wadood Chaudhary · All examples fictional

1. Agent and purpose. What it does, for whom, which tools, what it can change.

2. Failure modes. Read 30+ traces. Note the first thing wrong, as a cause. 4-7 modes with counts and severity.

3. Eval suite. Each mode → grader (code first), input, pass condition. Judges validated with kappa against adjudicated human labels.

4. Coverage. Router (including "no tool") · arguments · skill (result-set SQL, retrieval) · trajectory (mode plus order check) · outcome (end state).

5. Reliability and cost. 5+ trials. pass@1 with interval · pass^k · cost per success · p95 latency · n.

6. Security. Top 3 OWASP ASI risks · injection and benign cases · guardrail miss and false-alarm rates · tool-level controls · sandbox only.

7. Tracing and privacy. Content capture off · redact before export · retention · access · 201 CMR 17.00.

8. Online evaluation and monitoring. Cheap checks on all traffic; judges on a sample · alerts with thresholds · SLOs.

9. Release. CI gate · shadow · canary · rollback and kill switch · named sign-off.

10. Incidents and documentation. Postmortem trigger · failures become test cases · eval report · system card · NIST AI RMF Measure.