Evaluating and Operating Agentic AI Systems: Graders, Trajectories, Red-Teaming, and Production Monitoring

Course FAST-13 · 15 hours · Instructor: Wadood Chaudhary · Presented by 7formations

Session 1. Why agents fail and how to see it (3 hours)

- Cost of agent failure; public cases
- Agent anatomy: router, skills, memory, trajectory
- Vocabulary: task, trial, grader, trace, outcome; tests vs evals
- Error analysis: open and axial coding of traces
- The evaluation cycle (development and production loops)
- **Lab 1: Error analysis on fictional shipment-agent traces; failure taxonomy and evaluation plan.**

Session 2. Graders you can trust (3 hours)

- Code-based graders and outcome checks
- LLM-as-judge design; position, verbosity, and self-preference bias
- Validating judges: confusion matrix and Cohen's kappa; human annotation methods
- Golden datasets, coverage matrices, and synthetic data
- **Lab 2: Golden dataset, code graders, and a validated LLM judge; paired labeling with kappa.**

Session 3. Evaluating agent behavior end to end (3 hours)

- Router and tool-call evals (schema, AST, executable match, irrelevance)
- Skill evals: retrieval, SQL, code (sandboxed)
- Trajectory evals: match modes and convergence
- Planning, reflection, and outcome (end-state) grading
- Multi-turn, multi-agent, memory, computer-use, and MCP agents
- **Lab 3: Component and trajectory scorecard on a fictional quoting agent.**

Session 4. Reliability, cost, and security (3 hours)

- Non-determinism: multiple trials, pass@k vs pass^k
- Experiments, regression gates, cost and latency metrics
- Public agent benchmarks and their limits
- OWASP Top 10 for Agentic Applications; red-teaming; guardrails
- **Lab 4: Reliability, cost, and prompt-injection red-team report.**

Session 5. Operating agents in production (3 hours)

- Tracing with OpenTelemetry GenAI conventions; privacy controls
- Offline vs online evaluation with the same graders
- Monitoring, SLOs, alerts, canary and A/B releases, incident review
- Governance: eval reports, system cards, audit trails
- **Lab 5: Capstone Agent Evaluation and Operations Plan, presented and peer reviewed.**

Worksheets and deliverables

- W1 Failure taxonomy and evaluation plan
- W2 Golden dataset, graders, and judge-agreement report
- W3 Component and trajectory scorecard
- W4 Reliability, cost, and red-team report
- W5 Agent Evaluation and Operations Plan (capstone)
- Take-home: one-page agent evaluation and operations checklist

Attendance and completion

- Attendance is recorded at the start and end of every session (live virtual: platform attendance log plus an instructor roll call or poll; in-person: signed sheet).
- A trainee must attend 100% of instructional hours. A missed session may be made up by attending the same session in another scheduled cohort within 30 days; the make-up is logged the same way.
- A trainee must submit every required lab deliverable and score 'Meets' on the final applied task.
- Each completion record is tied to the trainee, the employer FEIN, and the course number, and is kept for the contract retention period.
- Incomplete trainees are not reported as completions and are not billed to anyone. FAST AI pays for completions only; employers and trainees are never charged.
- Optional asynchronous pre-work (under 10% of course time) is supplemental, is not counted as instructional hours, and never substitutes for a live session.

Assessment rubric

Each criterion is scored Not yet / Meets / Strong. Completion requires Meets on every criterion for the final deliverable (resubmission allowed within the course window).

Criterion	Meets standard
Failure-driven	Every eval targets a failure mode seen in traces, with a named grader
Grader quality	Code checks where possible; LLM judge agreement with humans reported

Criterion	Meets standard
Coverage	Router, tool-call, skill, trajectory, and outcome measures present
Reliability and cost	Multiple trials; pass^k and cost per success with sample size
Security	Injection cases run; guardrail misses and false alarms reported
Operations	Tracing, privacy controls, online evals, alerts, release gate specified