W4: Reliability, cost, and red-team report

Evaluating and Operating Agentic AI Systems: Graders, Trajectories, Red-Teaming, and Production Monitoring (FAST-13) · Lab 4: Reliability, cost, and red-team report

Name: _____ Date: _____

A table of v1 and v2 metrics with intervals, a one-paragraph ship / don't-ship recommendation, the attack results before and after the guardrail, miss and false-alarm rates, and OWASP mappings with a proposed control for each attack that got through.

Completion criteria. Both versions run with 5 trials; pass@1 with an interval, pass^5, cost per success, and p95 reported; all 24 emails run with and without the guardrail; both error rates reported; recommendation justified with numbers.

7formations · AI Training 2026 · Instructor: Wadood Chaudhary