Session 4 Handout: Reliability, Cost, and Security

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

Reliability

- **pass@1:** average success over all trials.
- **pass@k:** at least one success in k trials. Rises with k; fits tools where one good attempt is enough.
- **pass^k:** all k trials succeed. Falls with k; fits customer-facing agents. At 75% per trial, pass^3 ≈ 0.42.
- Report a **95% interval** (Wilson) with n. Trials of the same task are correlated, so the interval is optimistic; use a **paired bootstrap over tasks** to compare versions.

Cost and latency

Cost per successful task = total cost ÷ number of successes. Report p50 and **p95** latency.

Worked example (fictional OrderDesk, 20 tasks × 5 trials)

	v1	v2
Successes / runs	78 / 100	90 / 100
pass@1 (95% interval)	0.78 (0.69-0.85)	0.90 (0.83-0.94)
Total cost	$1.20	$1.30
Cost per success	≈ $0.0154	≈ $0.0144

v2 costs more per run but less per success. Confirm with a paired comparison and check pass^5 before shipping.

Benchmarks

Public agent benchmarks (SWE-bench Verified, tau-bench and tau2-bench, GAIA, WebArena, OSWorld, Terminal-Bench, BFCL) show general capability. They can saturate, leak into training data, or be gamed. Your own task suite decides.

Security: OWASP Top 10 for Agentic Applications (2026)

ASI01 Agent Goal Hijack · ASI02 Tool Misuse and Exploitation · ASI03 Identity and Privilege Abuse · ASI04 Agentic Supply Chain Vulnerabilities · ASI05 Unexpected Code Execution · ASI06 Memory and Context Poisoning · ASI07 Insecure Inter-Agent Communication · ASI08 Cascading Failures · ASI09 Human-Agent Trust Exploitation · ASI10 Rogue Agents.

Indirect prompt injection hides instructions in content the agent reads (emails, documents, web pages). Layer defenses: input classifiers, tool-level rules (approval over a limit), least privilege, and human approval for risky actions.

Measure guardrails both ways: miss rate (attacks that still succeed), false-alarm rate (benign inputs blocked), and added latency. Example: miss 2/12, false alarm 2/12.

Red-team only systems you own or are authorized to test.

Checklist

- ☐ Both versions run with 5 trials; intervals and pass^5 reported.
- ☐ Cost per success and p95 latency.
- ☐ 24 emails run with and without the guardrail; both error rates.
- ☐ Every successful attack mapped to an ASI category with a control (W4).