Session 3 Handout: Evaluating Agent Behavior End to End

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

Why this session matters

A right final answer can hide a wasteful or risky path, and a clean-looking path can still write the wrong row. Score each layer.

Component evals

Layer	Metric	Note
Router	Correct tool set; "no tool" on out-of-scope questions	Test both directions (should act, should not)
Arguments	Exact or normalized match	Normalize case and whitespace before comparing
SQL skill	**Result-set match** on a read-only connection	Different SQL text can be equally correct
Retrieval	Relevance, faithfulness, citation correctness	
Code skill	Runs in an **isolated** runner	Never exec() model-written code in the agent process

Trajectory match modes

Mode	Passes when
Strict	Same tools, same order, same count
Unordered	Same tools and counts, any order
Subset	Agent called nothing outside the reference
Superset	Agent called everything in the reference (extras allowed)

Superset checks presence, not order. To enforce "check_stock before create_quote", add an explicit order check. Prefer outcome grading unless the order is a real business rule.

Convergence: for paraphrases of one task, the average of S_optimal / S_run over completed runs. Steps 4, 4, 5, 8, 4, 6 give ≈ 0.83. It says nothing about failed runs, so report those separately.

Outcome grading

Compare the end state (the quote row) with the expected row. Example: QuoteDesk's path passed strict match but wrote qty 25 instead of 250. Only the outcome check (and an argument check) catches that.

Beyond single-turn

- **Multi-turn:** simulate users with a second model.
- **Multi-agent:** check that handoffs carry the needed context.

- **Memory:** check what gets stored and whether it's still right later.
- **Computer use:** verify the backend state, not the confirmation page.
- **MCP:** check the agent picks the right server's tool.

Checklist

- ☐ Router, argument, SQL, trajectory, outcome, and convergence rows filled (W3).
- ☐ Match mode chosen with a reason; order check added for the business rule.
- ☐ One "right outcome, wasteful path" case and one "clean path, wrong outcome" case found.