Session 1 Handout: Why Agents Fail and How to See It

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

Why this session matters

Agents pick tools, pass arguments, loop, and act. A wrong step early (a swapped digit in an ID) can produce a fluent, confident, wrong answer. You can't fix what you haven't named, so the course starts with reading traces and naming failures.

Key concepts

1. Agent anatomy

Part	What to evaluate
Router	Did it pick the right tool, or correctly pick none?
Tool call	Were the arguments right?
Skill	Retrieval, SQL, or code quality
Trajectory	The sequence of steps
Outcome	The end state (a row written, a ticket opened)

2. Vocabulary

- **Task:** one test case with inputs and success criteria.
- **Trial:** one attempt at a task. Agents vary, so run several.
- **Grader:** logic that scores part of a trial (code, LLM judge, or human).
- **Trace (transcript):** the full record of a trial: messages, tool calls, results.
- **Outcome:** the final state of the environment, which can differ from what the reply claims.

Tests vs. evals: unit tests check deterministic code. Evals measure model behavior as rates across many tasks and trials. You need both.

3. Error analysis

1. **Open coding:** for each trace, note the *first* thing that went wrong, as a cause ("digits swapped in shipment ID"), not a symptom ("bad answer").
2. **Axial coding:** group notes into 4 to 7 failure modes, each with a one-sentence definition.
3. **Count and rate severity**, then pick the top modes by count × severity.
4. **Propose a grader** for each: input and pass condition. Prefer code where possible.

Worked example (fictional ShipTrack traces)

Mode	Count / 40	Severity	First grader
Wrong ID extracted	7	High	Code: regex ID from user message = tool argument
Invented ETA (no get_eta call)	6	High	Code: ETA in reply requires a get_eta span
Handoff loop	4	Medium	Code: no assistant reply after open_handoff
Stale status shown as current	3	Medium	LLM judge with tool result
Off-topic answered	3	Low	LLM judge or keyword rule
OK	17		

Checklist

- ☐ 30+ traces open-coded with causes, not symptoms.
- ☐ 4 to 7 modes, each with a definition someone else could apply.
- ☐ Counts and severity recorded.
- ☐ Three graders proposed, at least one code-based.
- ☐ W1 submitted.