W1: Failure taxonomy and evaluation plan

Evaluating and Operating Agentic AI Systems: Graders, Trajectories, Red-Teaming, and Production Monitoring (FAST-13) · Lab 1: Error analysis on shipment-agent traces

Name: _____ Date: _____

Failure mode	Definition	Count (of N)	Severity	Proposed grader	Pass condition
e.g., Wrong ID extracted	Tool called with an ID that differs from the one the user gave	7 / 40	High	Code: compare tool `shipment_id` with the regex-extracted ID from the user message	IDs match exactly

Completion criteria. At least 30 traces coded; 4 to 7 defined modes with counts; three graders proposed, each with an input and a pass condition; at least one code-based grader.