W2: Golden dataset, graders, and judge-agreement report

Evaluating and Operating Agentic AI Systems: Graders, Trajectories, Red-Teaming, and Production Monitoring (FAST-13) · Lab 2: Golden dataset, graders, and a validated judge

Name: _____ Date: _____

Include the 30-item golden CSV, the grader pass rates, the human-human kappa, the judge-human kappa, the confusion matrix, and two sentences on where the judge disagreed and why.

Completion criteria. 30 golden items covering at least 10 matrix cells; both code graders run; judge run on all 30 replies; both kappas and the confusion matrix reported.

7formations · AI Training 2026 · Instructor: Wadood Chaudhary