Session 2 Handout: Graders You Can Trust

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

Why this session matters

A grader you haven't validated is just an opinion that runs fast. This session builds a golden dataset, code graders, and an LLM judge, then measures whether the judge agrees with people.

Grader types

Type	Best for	Watch out for
Code	Schema, formats, IDs, required tool calls, end-state checks	Brittle to valid variations
LLM judge	Faithfulness, tone, relevance, rubric criteria	Bias, cost, drift; must be validated
Human	Gold labels, expert calls, calibrating judges	Slow, costly, needs a rubric

Designing a judge

- **Binary labels** with clear definitions (faithful / unfaithful) are easier to validate than 1-10 scales.
- **Reason first, then label** on the last line, in a fixed format.
- Include a **fail example** in the prompt.
- **One judge per criterion.**
- Known biases: **position** (prefers the first answer: swap order and require consistency), **verbosity** (prefers longer answers), **self-preference** (favors its own model family).

Validating a judge

Percent agreement overstates quality when one label dominates. **Cohen's kappa** removes chance agreement: $\kappa = (p_o - p_e) / (1 - p_e)$.

Class guide: below 0.4 weak; 0.4-0.6 moderate; 0.6-0.8 substantial; above 0.8 strong. Also report the **confusion matrix** and the judge's recall on the failing class. For three or more labelers, use Fleiss' kappa or Krippendorff's alpha.

Worked example (fictional CoverageHelper)

20 replies; adjudicated human labels 14 faithful, 6 unfaithful. The judge agrees on 13 + 3, so $p_o = 0.80$. The judge says faithful 16 times, so $p_e = 0.70 \times 0.80 + 0.30 \times 0.20 = 0.62$. **$\kappa \approx 0.47$ (moderate)**: 80% agreement, but the judge caught only 3 of 6 unfaithful replies. Next step: tighten the rubric and re-measure on the same items.

Golden datasets

Build a coverage matrix (task type × difficulty × persona) and fill the cells that matter. Model-generated items are allowed only after a human reviews and edits each one. Include "should not act" cases, such as questions that need no ticket.

Checklist

- ☐ 30 golden items covering 10+ matrix cells.
- ☐ Both code graders run, with pass rates.
- ☐ Judge run on all replies.
- ☐ Human-human and judge-human kappa, plus the confusion matrix.
- ☐ W2 submitted.