Tutorial & Assignment Week 7 — Evals: Grading the Graders

Goals: Build an LLM grader and see how it scores on cases it has never seen!

Due: Submit the indicated portions on Gradescope by Sunday Nov. 8th at 11:59pm, and be ready to discuss with your TA during tutorial in Week 8.

Hi all! Last week you kept track of something that went wrong while working with your coding agent. This week we ask the question behind those examples: how do you actually know whether an AI is doing what you want?

In tutorial, everyone will build a grader in Eval Arena: you will label example answers and write instructions for an LLM that judges whether an AI assistant behaved responsibly. Every grader then gets scored on a hidden test set, and we will compare results as a group.

Assignment outline:

This document is intended to be started in tutorial and completed for homework if necessary. Work through it in the order given.

- Build and test a grader in Eval Arena
- Debrief the results as a group
- Submit your results and a short reflection

Part 1: Tutorial group work

Group Activity 1 — Eval Arena

Go to **eval-arena-production.up.railway.app** and enter the join code from the slides. You will be judging one behavior: *does this AI assistant handle the situation the way CS193T says a responsible AI assistant should?* The CS193T principles will be on the screen throughout.

Here's what you'll be doing:

1. **Label cases.** For each example answer, choose PASS or FAIL and write a one-line reason. Label at least 3; label as many as you like in about 20 minutes.
2. **Write your judging instructions.** Tell your grader how to decide.

3. **Check the grader preview, then submit.**

We will run every grader on the hidden test set and reveal the leaderboard together. Your score is balanced accuracy, so a grader that always says PASS (or always FAIL) scores 50%. A baseline grader, built with no examples and no instructions from you, appears on the board too. Your goal is to beat it.

Deliverable: You don't need to submit anything for this portion of the assignment.

Group Activity 2 — Debrief

As a group, we will look at the leaderboard, the hidden case that split the room the most, and what separated the graders that did well from the ones that did not.

Deliverable, copied and pasted into Gradescope:

1. Your score

Part 2: Reflection — How did it go?

Write one short paragraph, approximately 4–6 sentences, about how Eval Arena went for you. How did your grader do compared to the baseline, and what do you think made the difference? Did anything surprise you when the results came in?

Deliverable: Copy and paste your reflection into Gradescope.

Submission

You should have two items to submit to Gradescope. Be ready to discuss these with your TA next week:

1. Your Eval Arena score
2. Your reflection!

Once you're done, use this time to work on your final project!