MS&E 203 · LECTURE 3 · EVAL WORKFLOW CARD

Your eval workflow card

Six steps, one card. Fill in steps 1–4 in class; steps 5–6 feed your Week 4 proposal. The running example in lecture is a clinical scribe; swap in your own product.

Name _____ **Product** _____

1 Start from the scenario

Evals are only meaningful relative to a job. Name the job concretely, and name everyone who uses the output.

In → out What goes in, and what comes out?

Consumers Who uses the output, and what does each one need? They rarely want the same thing.

Cost of failure

Baseline How the user does this today

Decision

```
We will ____________ if _________________________, compared with ____________.
```

2 Define the dimensions you care about

Several dimensions, scored separately, each with a pass/fail example and a severity tier. Importance lives in the decision rule, not a blended score.

DIMENSION	QUESTION IT ASKS	PASS EXAMPLE	FAIL EXAMPLE	TIER

Tier: C = critical (blocks the output) · M = major (revise first) · m = minor (informs iteration)

After reading 20–30 real outputs Which dimension will you add or split?

Proxy and outcome If you track a proxy (edits, clicks), which outcome will you track next to it?

3 Build an initial set of cases

Span the space, not just the happy path. Real data beats invented data; have an expert label a gold set.

Common cases Where will they come from, and how many?

Corner cases (2+)

Known failure modes (2+)

Collections; dev/test split Estimate vs. hunt? Held out?

Groups to report separately Who could it serve worse?

One case, written out Input · expected output · how you will check it

4 Evaluate in layers

Cheapest first. Calibrate the judge before you trust it. Findings move to cheaper layers over time.

Layer 1 · automated check One rule you could check in code on every output

Layer 2 · LLM judge Which dimensions? How many human-labeled cases to calibrate it?

Layer 3 · expert review Who reviews (spanning your consumers)? Flagged cases plus a random sample?

Read the results Threshold you will set before you look

Advance if

Revise if

Collect more evidence if

5 Improve your product · Week 4

What users change is labeled data for free. Validate it against an expert sample so silent errors don't read as wins.

Edits or behaviors you will capture

6 Close the loop · Week 4

Match the fix to the diagnosis and start cheap: prompt → template → pipeline → fine-tuning. Harden every fix into a check.

First fix to try, and the check that keeps it from coming back

EXIT TICKET + ATTENDANCE

"The decision my eval should support is …" "The first check I will automate is …"

Submit both sentences at forms.gle/1FGw6vk2hk4NDpvD9, or scan the code. This also records attendance.