THE BUILDER'S STACK · THE PRACTICE · EVALS

# The Quality Bar

Your product's definition of good, made operational: the bar, the cases, the graders, the gate every change must clear, and the conveyor that keeps it honest. A sheet you run, not a sheet you file. Fill it once, then keep it beside the product and re-run it weekly.

**PRODUCT / FEATURE**

**OWNED BY**

**DATE**

## 1 The job and the bar

*Good derives from the job users hire the product for. Name the job, the one load-bearing dimension, and five statements written so two people score them the same way.*

**THE JOB USERS HIRE THIS PRODUCT FOR**

**THE LOAD-BEARING DIMENSION**

**BAR STATEMENT 1**

☐ Must never fail this ☐ Must always pass ☐ Target to raise

**BAR STATEMENT 2**

☐ Must never fail this ☐ Must always pass ☐ Target to raise

**BAR STATEMENT 3**

☐ Must never fail this ☐ Must always pass ☐ Target to raise

**BAR STATEMENT 4**

☐ Must never fail this ☐ Must always pass ☐ Target to raise

**BAR STATEMENT 5**

☐ Must never fail this ☐ Must always pass ☐ Target to raise

## 2 The case ledger (starter)

*Real usage first, synthetic second, adversarial always. Every case carries its expectation. The golden set is the regression backbone; production grows the living set.*

**CASE 1**

**INPUT (SUMMARY)**

**SLICE**

**A PASS LOOKS LIKE**

**CASE 2**

**INPUT (SUMMARY)**

**SLICE**

**A PASS LOOKS LIKE**

**CASE 3**

**INPUT (SUMMARY)**

**SLICE**

**A PASS LOOKS LIKE**

**CASE 4**

**INPUT (SUMMARY)**

**SLICE**

**A PASS LOOKS LIKE**

**CASE 5**

**INPUT (SUMMARY)**

**SLICE**

**A PASS LOOKS LIKE**

**CASE 6**

**INPUT (SUMMARY)**

**SLICE**

**A PASS LOOKS LIKE**

## 3 Graders

*Deterministic checks first: free and perfectly consistent. Judges only with an anchored rubric, order-swapped, spot-audited. Humans calibrate; rater disagreement means the rubric is broken.*

**DETERMINISTIC CHECKS (SCHEMA, CONTAINS, BANNED CONTENT, BOUNDS)**

**JUDGE RUBRIC (ANCHORED WITH EXAMPLES) AND WHAT IT SCORES**

- [ ] Judge runs pairwise with answer order swapped
- [ ] A sample of judge verdicts is human-audited every run

**HUMAN GOLD LAYER: WHO READS, HOW MANY, HOW OFTEN**

## 4 The regression gate

*Every behavior-shifting change runs the set before it merges. Yesterday's bar is the floor. A model upgrade is a change like any other.*

**FAST SUBSET, RUN ON EVERY CHANGE**

**FULL SET, RUN ON**

**THE FLOOR (YESTERDAY'S BAR)**

- [ ] Model version pinned; full re-run before any upgrade ships

## 5 The conveyor

*Production keeps the eval honest: instrument the signals, review on a schedule, and turn every failure into a case before the week ends.*

**SIGNALS INSTRUMENTED (EXPLICIT AND IMPLICIT)**

**TRANSCRIPT REVIEW CADENCE**

**FAILURE BECOMES A CASE WITHIN**

## 6 The Goodhart pair

*When a measure becomes a target, it stops measuring. Name the gameable metric, give it a counterweight, and keep cases the tuning never sees.*

**THE MOST GAMEABLE METRIC**

**ITS COUNTER-METRIC**

☐ Hold-out cases exist that tuning never touches

**SCHEDULED HUMAN NORTH-STAR READ (WHO, WHEN)**

## 7 After the first run

*Record the floor while the run is fresh, and the first thing the bar caught that the demo never would have.*

**FIRST RUN DATE**

**THE FLOOR WE RECORDED**

**THE FIRST THING THE BAR CAUGHT**

**KEEP FOR NEXT TIME**