Agent evaluation checklist

Datasets, graders, trajectory and state checks, and release gates.

QA | Developer | Product

DATASET

- About 50 real or realistic cases to start
- Common, ambiguous, out of scope, hostile, injection and policy-edge cases
- Tag by scenario; keep a holdout
- Every production failure becomes a case

WHAT TO GRADE

Dimension	Check with
Final response	Code, then a calibrated judge
Trajectory	Expected path or rules
Tool use	Code: tool, arguments, call count
State changes	Code: query the system after the run
Whole conversation	Judge over the transcript plus human samples

JUDGES

- One criterion each, written as pass or fail
- Give the evidence: policy, tool results, transcript
- Label 50 to 100 cases with people first and measure agreement
- Randomize order and test for length bias
- Re-calibrate when the judge, rubric or agent changes

SIMULATED USERS

- Personas with goals and constraints, including difficult ones
- Limit turns and record everything
- Compare with real transcripts; do not trust them alone

RELEASE GATE

- Fail on any safety case
- Fail on a drop beyond a threshold set from your own baseline
- Fail when a fixed bug returns
- Run offline before release and sample online after