

Classic Training

WEEKLY DIGEST

17-23 August 2026



This week

17-23 August 2026

Available archived editions, in their original date order.

MONDAY / 17 August 2026

Agent Evals 101

[Open original online](#)

3

AGENT EVALS 101

TEST BEFORE YOU TRUST



TEST. MEASURE. IMPROVE. `</>`

Hi! I'm **Elly**. Today we'll learn how to **test** an AI workflow before **trusting** it.

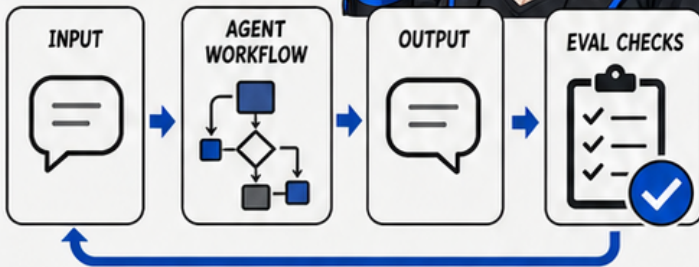
Good agents are tested — not guessed.

MEASURE DON'T GUESS

`</>`

WHAT IS AN EVAL?

An eval is a **repeatable test** of an AI system's behavior.



Run it. Check it. Learn. **Repeat.**

WHY DO WE NEED IT?

One good demo proves almost nothing. Real cases reveal what **works**— and what **breaks**.



GOOD DEMO



- ⚠ One happy case
- ⚠ Easy input
- ⚠ Cherry-picked

Looks good...

VS.

REAL CASES



- Many scenarios
- ✓ Edge cases
- ✓ Real-world data

Tells the truth.



TODAY'S GOAL:
Turn vague quality into **measurable** checks.



Define success



Test cases



Scores



Better behavior

FROM GOAL TO TEST CASES

PAGE 2



1

DEFINE THE TASK

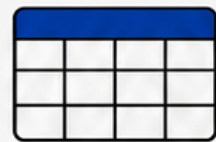
What should the agent do?



2

COLLECT CASES

Use real, typical inputs.



3

ADD EDGE CASES

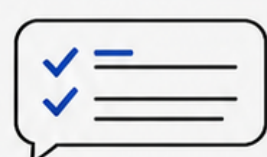
Include ambiguity and failure risks.



4

WRITE EXPECTATIONS

Describe acceptable behavior.



5

RUN AND SCORE

Measure results consistently.



KEY TAKEAWAY

Good test cases represent *the work*—not just the happy path.



Real cases



Edge cases



Reliable learning



THE 4 PARTS OF A USEFUL EVAL

PAGE 3



1. DATASET

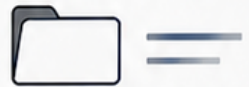
Inputs that represent real work.



Realistic examples



Balanced coverage



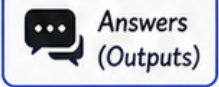
Organized & versioned

2. RUNNER

The workflow that produces answers.



Agent Workflow



Answers (Outputs)

3. GRADER

Rules or reviewers that score behavior.



Correctness
Right vs. wrong



Completeness
Did it cover what matters?



Safety
Followed rules and avoided harm?



Tone / Quality
Clear, helpful, and well-formed?

4. THRESHOLD

The score required to pass.



Set the bar for quality.

- ☒ Too low = poor quality slips through
- ☒ Too high = hard to ship anything

★ Pick a threshold that matches the risk.

GUARDRAILS



Use realistic cases

Reflect real tasks and edge cases.



Protect private data

Mask or remove sensitive information.



Keep scoring stable

Clear rules and consistent graders.

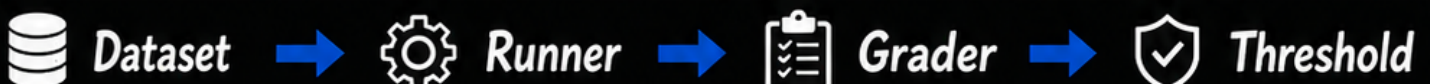


Inspect failures

Learn from misses and improve.



A single average can hide a **dangerous** failure.



A FAILED SUPPORT AGENT

TEST, FIX, RETRY

PAGE 4



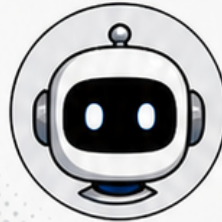
1. TEST CASE



Can I return a used item after 60 days?



2. BAD ANSWER



Yes—just send it back.



3. SCORE



Policy accuracy: 0/1



Evidence: 0/1



Escalation: 0/1



4. DIAGNOSE



It *guessed*.
It did not check the policy or flag the exception.



5. FIX



Retrieve the policy



State the 30-day limit



Escalate exceptions

6. RETRY



Standard returns are within *30 days*.
This case needs *human review*.

PASS: 3/3



BUILD, TEST, LEARN

THE EVAL LOOP

EVALUATE.
IMPROVE.
DELIVER
BETTER.



Evals make improvement **visible**. Run them whenever the workflow changes.



PRACTICAL CHECKLIST



☒ Define success



☒ Cover real and edge cases



☒ Use clear graders



☒ Inspect failures



☒ Rerun after changes

REMEMBER



An eval **does not** prove perfection.

It measures performance on the cases **you chose**.



≠



FINAL TIP FROM ELLY

Test the behavior you **need**—not the answer you **hope** for.



Measure **honestly**. Improve **deliberately**. Ship with **evidence**.

