

A STORYBOARD IN 6 FRAMES

What counts as good, and how to check it

*better at what,
exactly?*



engrave the blank first

Evaluation compares behaviour against an expectation written down before the run. Every faster, cheaper check approximates that written judgment.

Check a property, not “quality”.



*moves without saying
which part moved*

Does it parse?

*fields present,
types right*

Is it supported?

*by the source
it cites*

Does it decline?

*what it's
meant to*

Right register?

*the one the product
committed to*

Each property fails differently and is checked by different means.
One blended score hides which of them moved.

Four checks, one judgment.

1

Deterministic assertions

exact, cheap, narrow

⋮

2

Heuristics

a tripwire, not a definition of correct

⋮

3

Model-based graders

judgment at volume, same probabilistic kind

⋮

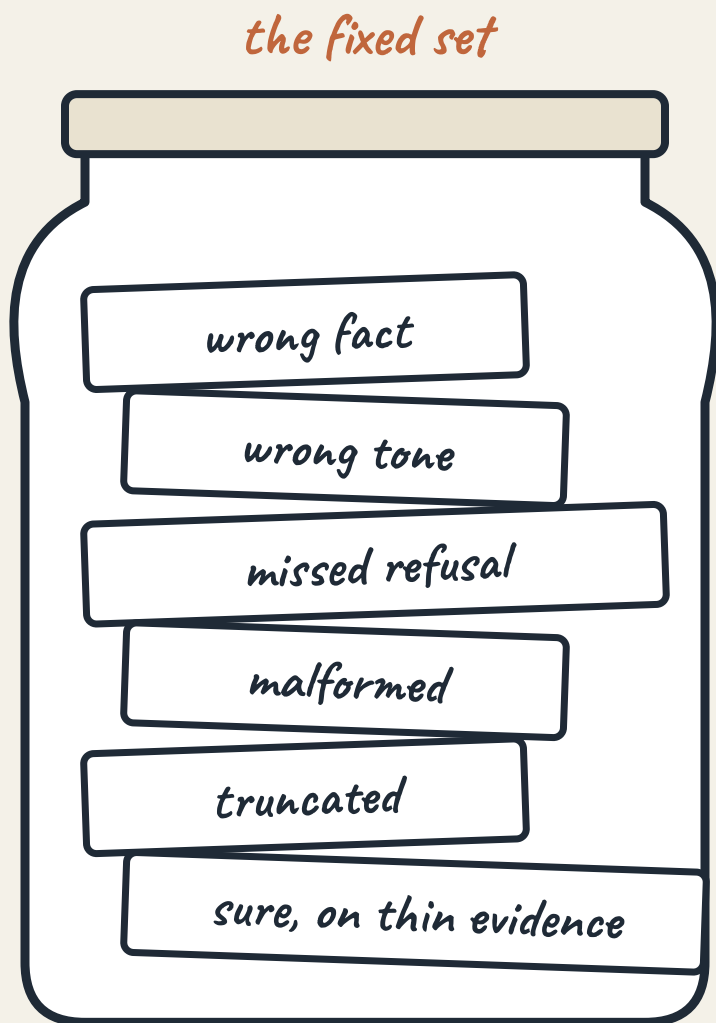
4

Human review

the one the other three stand in for

Every cheaper check approximates a judgment someone made first. The judgment doesn't disappear when it's automated; it gets encoded.

Keep the bad outputs.



change something

rerun the same set

see what moved

only your change moved

~~fresh samples~~ → **the same set**

It isn't an estimate of production traffic. It's failures that happened at least once, kept so their return gets noticed instead of reported by a user.

A number needs a decision.

1

What is measured

the property, named

⋮

2

On what population

whose runs, which cases

⋮

3

Over what period

from when, to when

⋮

4

How it was produced

the method, not only the result

⋮

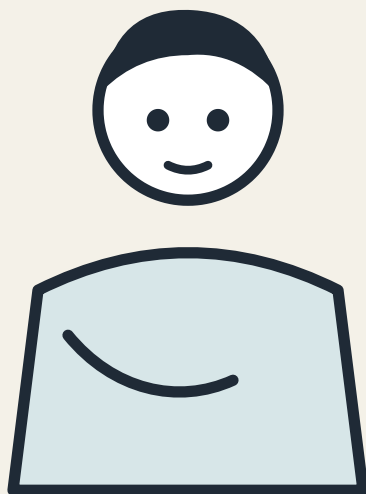
5

Which decision changes

if none would, drop the metric

A figure quoted without its population and method hardens into a fact, and the conditions it was true under stop travelling with it.

Write down what “worse” looks like.



*worse =
a confident answer
with no source*

before the change, not after

Before the next change, describe behaviour you wouldn't accept, concretely enough that someone else could look at an output and agree it happened. Everything else checks that.

Read the whole explanation

[pruningmypothos.com/systems/
evaluation-is-a-human-problem/](https://pruningmypothos.com/systems/evaluation-is-a-human-problem/)

