

**“*AI Eval*” really means
“*Estimation of Random Variables*”**

Evals in AI

AI System under test
(e.g. ML classifier)



Objectives (goal of evals)



And Metrics

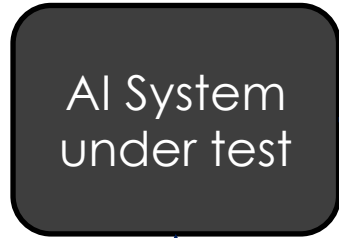
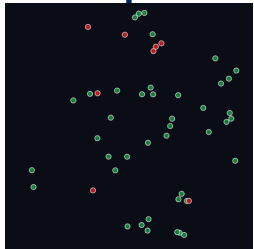
Scorecard	Value
Overall Accuracy	
Recall for class X	
Cost	
Latency	
...	

AI System
under test

We care about (future)
production value

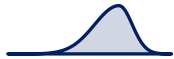
Scorecard	Value
Overall Accuracy	
Recall for class X	
Cost	
Latency	
...	

And we can only **compute** how the system performs on those datapoints



Scorecard	Value
Overall Accuracy	83
Recall for class X	91
Cost	0.02
Latency	35ms
...	

Based on that, we need to inform our belief

Overall Accuracy	
Recall for class X	
Cost	
Latency	
...	

But all we have is a test dataset

To estimate probability distributions of metrics we
need to run experiments

We run experiments to inform our belief

Bad experiments => False belief =>
Wasted improvement loops + Bad decisions

How to design good experiments?

Experiment Design

System Under
Test

Objectives and
metrics

Datasets

Eval workflow and assessment of quality

Reporting

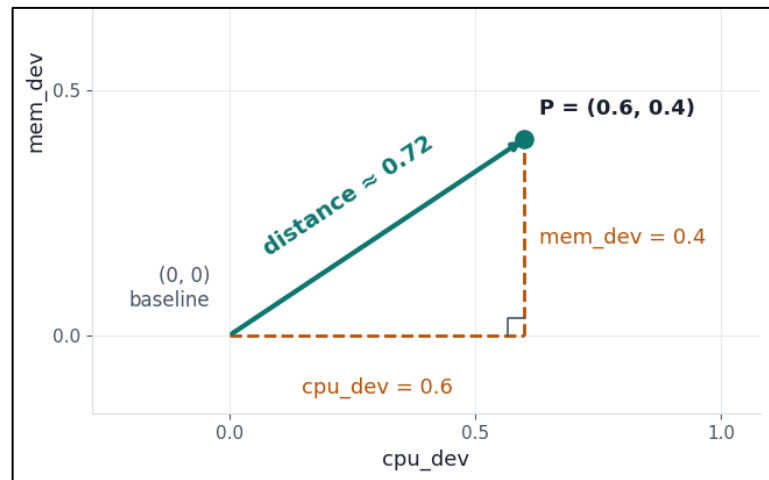
E1: Host health anomaly score in monitoring

```
def deviation_score(cpu_dev: float, mem_dev: float) -> float:

    if not 0.0 <= cpu_dev <= 1.0:
        raise ValueError(f"cpu_dev must be in [0, 1], got {cpu_dev}")

    if not 0.0 <= mem_dev <= 1.0:
        raise ValueError(f"mem_dev must be in [0, 1], got {mem_dev}")

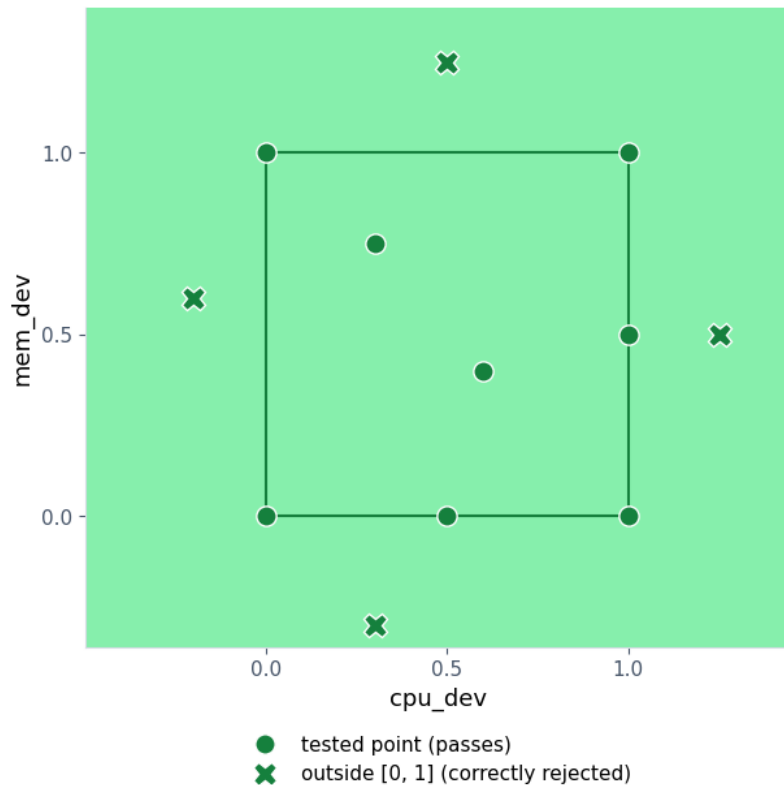
    return (cpu_dev**2 + mem_dev**2) ** 0.5
```



For simple functions, a somewhat easier task

E1

Anomaly score



E2: ML classifier for routing tickets to groups

A *classification* task: classify into ONE of these 5 classes:
account, order, refund, payment, delivery.

Training time

input

Thousands of customer requests (with solutions)

Classification
ML trainer

output

Classifier

Inference time

input

how do I get a refund for my last payment?

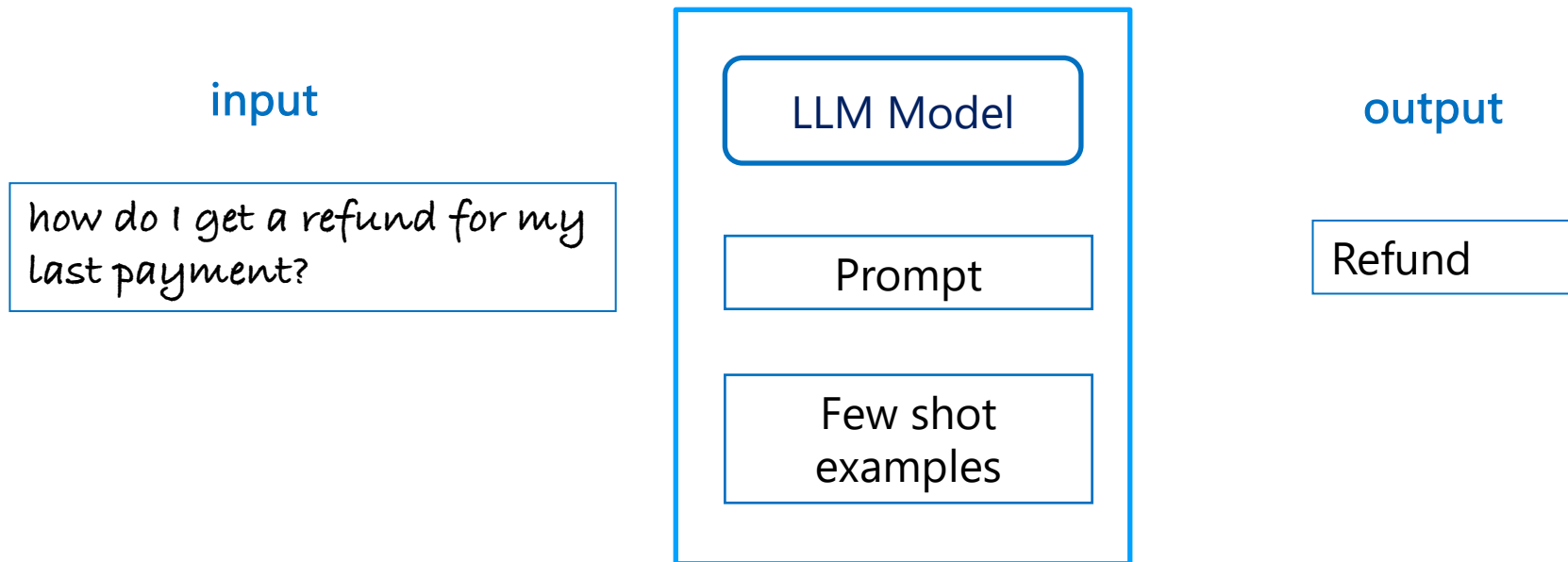
Classifier

output

Refund

E3: GenAI, few shot classifier for routing tickets to groups

A *classification* task: classify into ONE of these 5 classes:
account, order, refund, payment, delivery.



	ML Classifier	Gen AI Classifier
System under test	ML Training and inference (E2E) Variations: Different n. of params, different regularization ,..	
Objectives and metrics	Can we reach >90% accuracy? For which domains / kind of tickets?	
Datasets	Training: 100K cases from customers X,Y,Z Inference/test: 1000 cases from X,Y,Z,A,B	

Eval is System – and is the foundational building block for AI development

*Quality in AI systems depends on our ability to identify what to improve and how.
Bad eval = bad system and wasted iterations
LLMs work in the same way*

AI System X

*This is a system. Subject to design, development, reviews, versioning, approvals, ...
All the practices we have learned over the years*

Eval for AI
System X

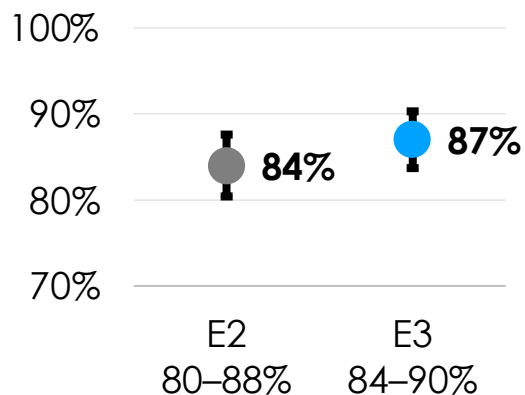
Also a system. It should be subject to design, development, reviews, versioning, approvals, ...

And evals should be repeatable and observable

Reporting

Accuracy, n = 400

Illustrative data



E3 +3 points, interval -1 to +7: no proven winner

Confusion matrix

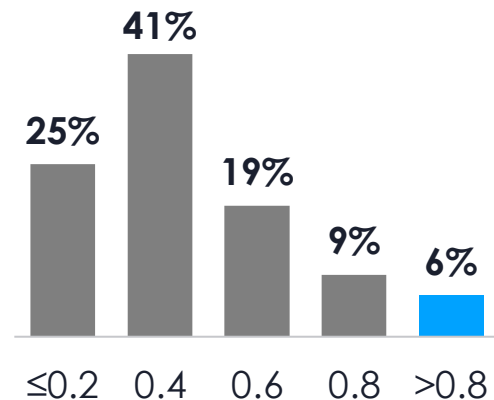
E3, true × predicted

	1	2	3	4	5
1	71	3	4	1	1
2	2	70	1	1	6
3	4	0	72	2	2
4	1	1	5	68	5
5	4	5	2	2	67

Class 5 absorbs the most errors: 14 of 52

Latency, seconds

E3, % of requests



6% of requests take over 0.8 s

Deadly reporting:
“The Slide that Killed 7 people”



Review of Test Data Indicates Conservatism for Tile Penetration

- The existing SOFI on tile test data used to create Crater was reviewed along with STS-87 Southwest Research data
 - Crater overpredicted penetration of tile coating significantly
 - Initial penetration to be described by normal velocity
 - Varies with volume/mass of projectile (e.g. 200ft/sec for 3cu. Ln)
 - Significant energy is required for the softer SOFI particle to penetrate the relatively hard tile coating
 - Test results do show that it is possible at sufficient mass and velocity
 - Conversely, once tile is penetrated SOFI can cause significant damage
 - Minor variations in total energy (above penetration level) can cause significant tile damage
 - Flight condition is significantly outside of test database
 - Volume of ramp is 1920cu in vs 3 cu in for test

Thanks :)

<https://reliableai.github.io>