MACHINE LEARNING RELIABILITY LAB

Model Reliability Case File

Your evidence for deciding whether a model should be approved, revised, restricted, or rejected

NAME	CLASS / PERIOD
_____	_____
MODEL OR PROJECT	**TEAMMATE(S)**
_____	_____

THE QUESTION

How do we know whether a model works well enough for the people who will depend on it?

HOW THE CASE FILE WORKS

1	2	3	4	5
Plan fair tests	**Find failure patterns**	**Audit representation**	**Compare mistake costs**	**Make a release decision**

STARTING PREDICTION

Before testing, what do you expect this model to do well? What do you expect might make it fail?

_____

_____

_____

CHECK BEFORE MOVING ON

☐ I will record evidence, not guesses

☐ I will protect privacy while collecting test data

DAY 1 | Plan a Fair Test

Working in a demo is not the same as working reliably.

OBJECTIVE **I can design a test suite that changes one condition at a time.**

INTENDED USE
What job is this model supposed to do? Who would use or depend on it?

TEST CONDITIONS

Choose at least three conditions. Include a normal case, an edge case, and an intentionally difficult case.

SELECT CONDITIONS

☐ lighting	☐ distance	☐ camera angle	☐ background
☐ different person/object	☐ partly blocked	☐ movement	☐ your idea

12-TEST PLAN

#	What I will show	Condition changed	My prediction	Why this is a fair test
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				

PREDICTION CLAIM
I predict the model will be least reliable when _____ because _____.

DAY 2 | Stress-Test the Model

Run the plan, record every result, and look for a pattern.

OBJECTIVE **I can use test evidence to identify a repeated failure condition.**

RESULT CODES

C Correct prediction	W Wrong prediction	U Uncertain or low confidence

TEST LOG

#	Actual label	Predicted label	Changed condition	C/W/U	Evidence or note
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					

Correct	Wrong	Uncertain
_____ / 12	_____ / 12	_____ / 12

FAILURE-PATTERN CLAIM

The model was least reliable when _____. In _____ of _____ tests, it _____.

DAY 3 | Audit Representation

More data is not automatically better data.

OBJECTIVE **I can connect uneven performance to a specific gap in the model's evidence.**

REPRESENTATION MAP

Condition or group	Well represented	Weakly represented	Not tested
1. _____	☐ Evidence: _____	☐ Evidence: _____	☐ Evidence: _____
2. _____	☐ Evidence: _____	☐ Evidence: _____	☐ Evidence: _____
3. _____	☐ Evidence: _____	☐ Evidence: _____	☐ Evidence: _____

EVIDENCE GAP

What important condition, person, object, or environment has weak or missing evidence?

_____

_____

_____

TARGETED REPAIR

Name the exact examples you would add, remove, rebalance, or test. Do not write only “add more data.”

_____

_____

_____

WHY THE GAP MATTERS

Who or what could be affected if the model performs unevenly in this condition?

_____

_____

_____

CHECK BEFORE MOVING ON

☐ I named a specific evidence gap ☐ My repair names exact examples ☐ My claim uses test evidence

DAY 4 | Compare the Cost of Mistakes

The more serious error depends on what the model is being asked to do.

OBJECTIVE **I can compare false alarms and misses for a specific use.**

FALSE ALARM The model says something is present when it is not.	MISS The model fails to detect something that is present.
Example from today's scenario: _____ _____	Example from today's scenario: _____ _____

SCENARIO AND INTENDED USE

What is being classified? Who would act on the result?

LABELING POLICY

Sample case	Our label + reason	Where people might disagree

ERROR-COST DECISION

For this use, which is more serious: a false alarm or a miss? Explain the consequence and who would be affected.

DAY 5 | Complete the Model Card

State what the evidence supports and what remains unknown.

OBJECTIVE **I can document a model's intended use, evidence, limitations, and next improvement.**

MODEL NAME
Name the model or project.

INTENDED USE + USERS
What job is it for? Who would use or depend on it?

TEST EVIDENCE
How many tests did you run? Where did it perform best? Where was it least reliable?

LIMITATIONS + NON-USE
What was not tested? What should this model not be trusted to decide?

REPRESENTATION CONCERN
Who, what, or which conditions have weak evidence?

NEXT IMPROVEMENT
Name the exact data, test, rule, or design change the team should make next.

FINAL DECISION | Approve, Revise, Restrict, or Reject

A responsible engineering decision is supported by evidence, not optimism.

CHOOSE ONE DECISION

☐ APPROVE	☐ REVISE	☐ RESTRICT	☐ REJECT
Evidence supports the named use.	Improve and retest before use.	Use only under named limits.	Risk or evidence gap is too large.

EVIDENCE-BASED RECOMMENDATION

We recommend _____ this model for _____ because our tests showed _____.

BOUNDARY

This model should NOT be used to _____ because _____.

NEXT ACTION

Before broader use, the team should _____ and then test _____.

60-SECOND BRIEF

1. Intended use	This model is designed to _____.
2. Strongest evidence	Our strongest evidence is _____.
3. Key limitation	The biggest limitation is _____.
4. Decision	We recommend _____ because _____.

FINAL SUCCESS CRITERIA

☐ I used numbers or specific test results	☐ I named at least one limitation
☐ My decision matches my evidence	☐ I explained the next improvement