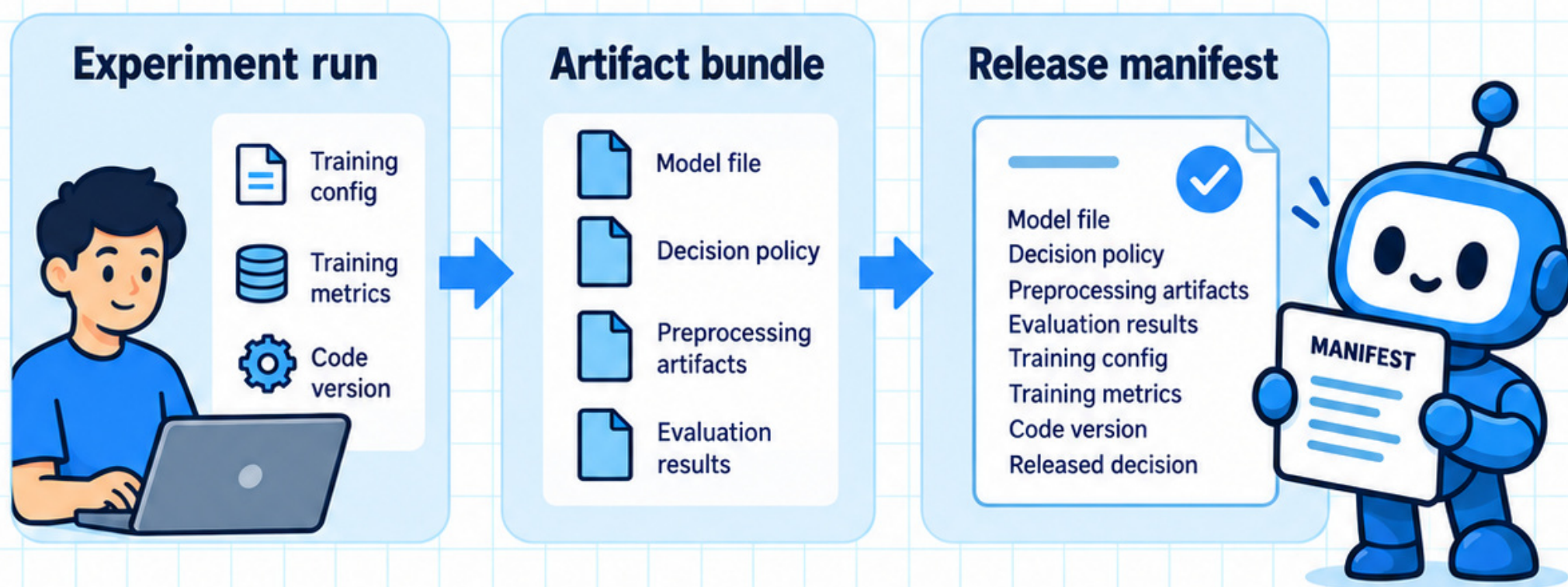


CAN YOU RECONSTRUCT YESTERDAY'S MODEL?

Track experiments and artifacts

A useful result needs more than a model file.
Save the evidence that connects training,
evaluation and the released decision.

Today: build a release record you can inspect and replay.



AN EXPERIMENT CONTAINS COMPARABLE RUNS

An experiment groups a question, such as which baseline works for our review queue.

Experiment

Which baseline works for our review queue?

Run A

Run B



A run records one execution: inputs, settings, outputs and measured results.

Run A

-  Inputs
-  Settings
-  Outputs
-  Measured results

Run B

-  Inputs
-  Settings
-  Outputs
-  Measured results

Give each run a stable ID. A memorable filename is not a complete record.



A memorable filename is not a complete record.



SEPARATE SETTINGS FROM OBSERVATIONS

PARAMETERS

Parameters describe choices: feature set, regularization and random seed.



- ✓ feature set
- ✓ regularization
- ✓ random seed

METRICS

Metrics describe measurements: validation recall, review volume and training duration.



validation recall



review volume



training duration

ARTIFACTS

Artifacts are saved outputs: fitted pipeline, evaluation tables and plots.



fitted pipeline



evaluation tables



plots



Record how each metric was calculated.



HOLD THE COMPARISON CONTRACT FIXED

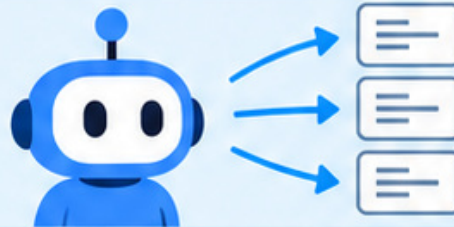


Toy validation cohort: the same eight customers from Day 11. Both candidates receive three reviews.

Same cohort
(the eight customers from Day 11)

Same capacity
(three reviews each)

Recency rule:



precision 1/3;
recall 1/4.

Learned-score ranking:



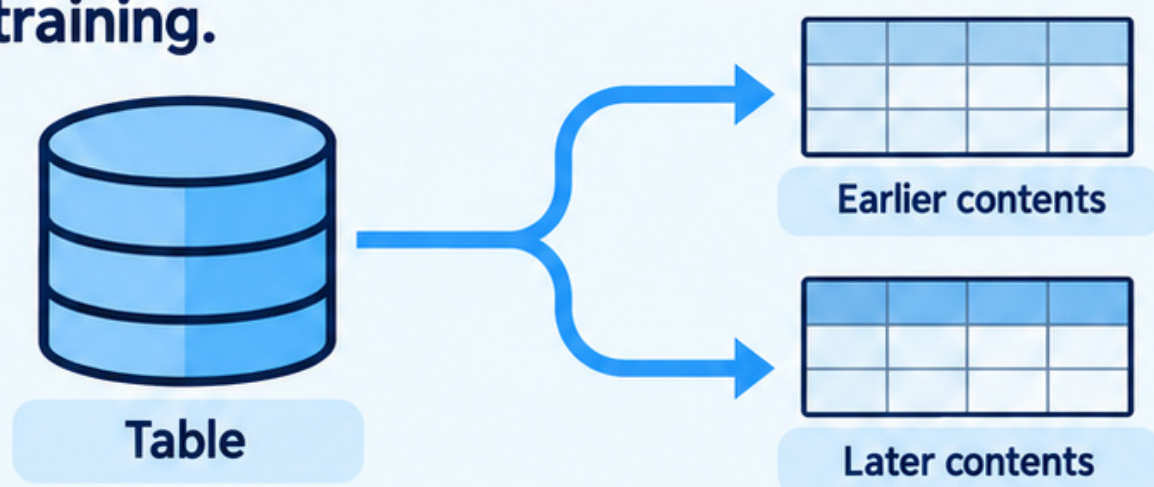
precision 2/3;
recall 2/4.



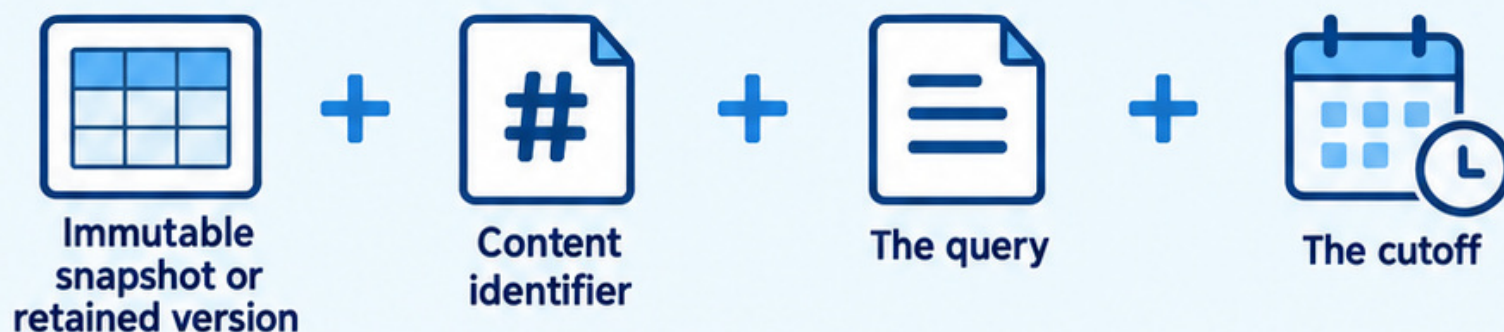
A changed cohort or budget needs a separate comparison. These figures are teaching data.

A DATA LINK CAN POINT TO CHANGING CONTENTS

Saving a table name or file URL does not freeze the rows used for training.



Record an immutable snapshot or retained version, its content identifier, the query and cutoff.



Keep the actual snapshot available under the data retention policy. A hash cannot reconstruct missing data.



SAVE THE BOUNDARIES THAT MADE THE RESULT VALID

Record the training, validation and final-test membership or reproducible split rule.



Include time cutoffs, label availability and any group separation.



time cutoffs



label availability



group rules

A seed alone cannot recover a split if the input rows or ordering changed.



same seed

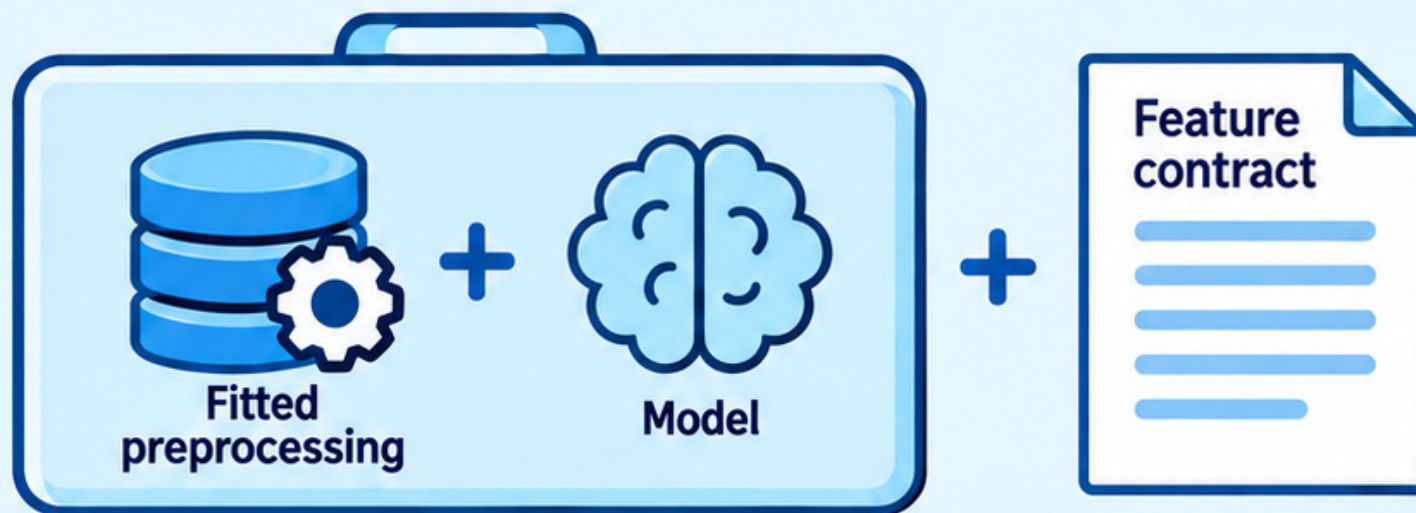


different input rows or ordering



PACKAGE THE FITTED TRANSFORMATION WITH THE MODEL

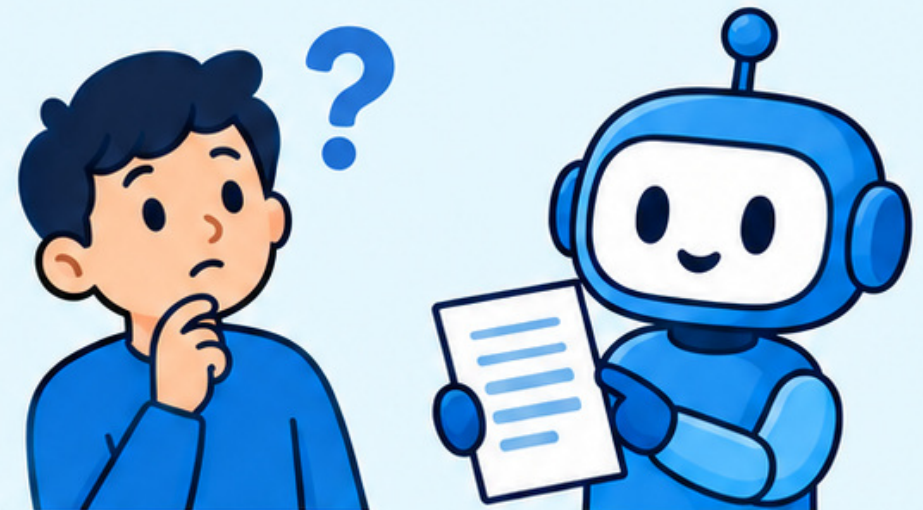
The release needs the fitted preprocessing and model, plus the expected feature contract.



Preserve column names, units, missing-value rules and categorical meanings.

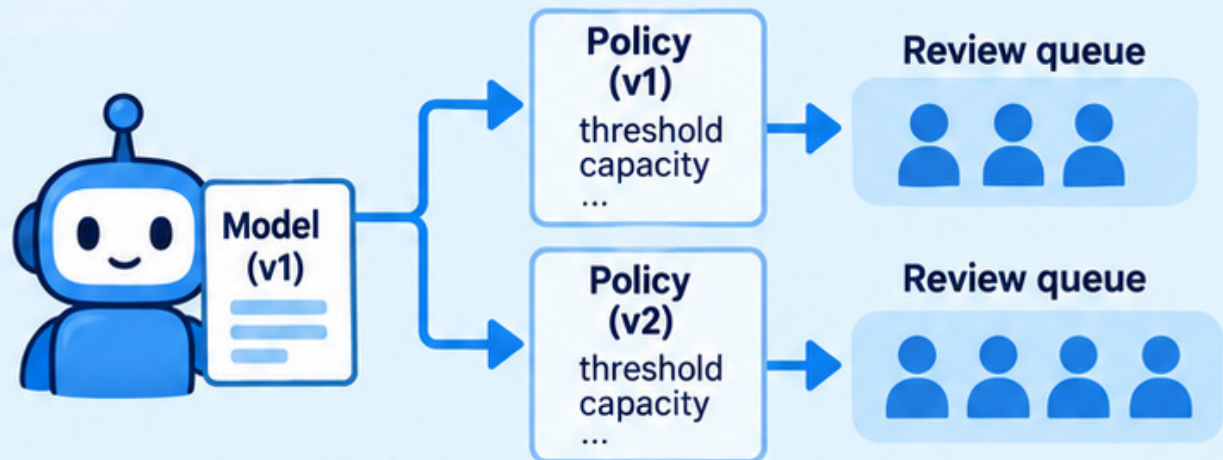
Column name	Unit	Missing-value rule	Categorical meaning

A correctly shaped input can still use the wrong unit or meaning.

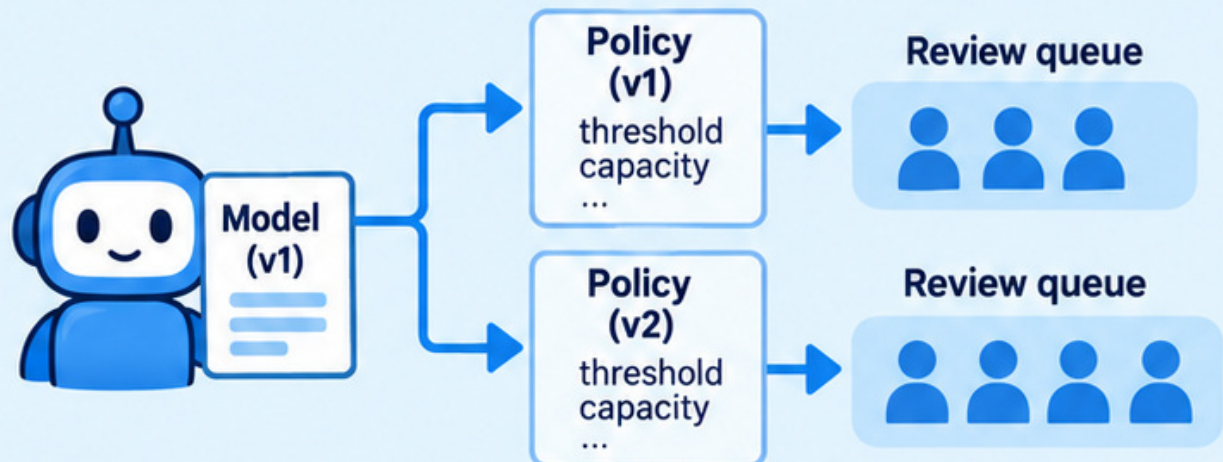


THE DECISION POLICY HAS A VERSION TOO

Our application selects $\text{score} \geq \text{threshold}$ and applies its review-capacity policy.



Changing the threshold can change behavior even when model bytes stay identical.



Record threshold, capacity, tie-breaking and overflow behavior in the release.



WRITE ONE RELEASE MANIFEST

Link these records:

- Run ID and source-code revision
- Data snapshot and split definition
- Label and feature definitions
- Fitted artifact and dependency versions
- Decision-policy version
- Evaluation scope and acceptance result



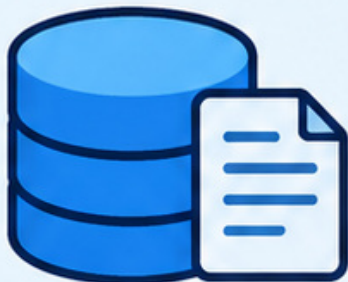
The manifest connects evidence to behavior.

USE MLFLOW TO ORGANIZE THE EVIDENCE

MLflow Tracking records runs, parameters, metrics and artifacts.



Dataset metadata helps identify run inputs. It does not by itself retain every source row.



Dataset metadata

Helps identify run inputs.

It does not by itself retain every source row.

The application still owns data retention, valid comparisons and release acceptance rules.



- ✓ Data retention
- ✓ Valid comparisons
- ✓ Release acceptance rules

AN INPUT SIGNATURE CHECKS ONLY PART OF THE CONTRACT

An MLflow model signature can describe expected input and output types or shapes.

Model Signature

```
purchase_amount: float  
user_id: string  
output: float
```



It cannot prove that a numeric purchase amount is in the intended currency or that a feature was available on time.



Add semantic and time-boundary checks beside the signature.



Type/shape
check
(from signature)



Meaning/unit
check

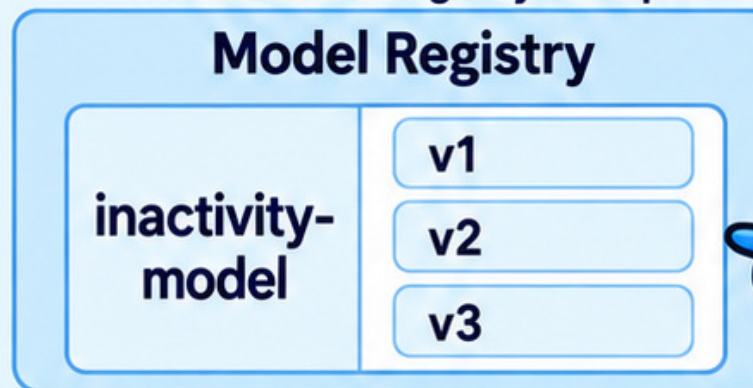


Time
availability
check

RESOLVE A NAME TO A CONCRETE VERSION

A model registry organizes named models and their versions.

Illustrative registry example



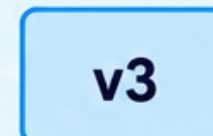
An alias is a movable reference. It can point to a different version later.



Resolve the alias when creating a release or batch, then record and use the concrete version and artifact identity.



Resolve

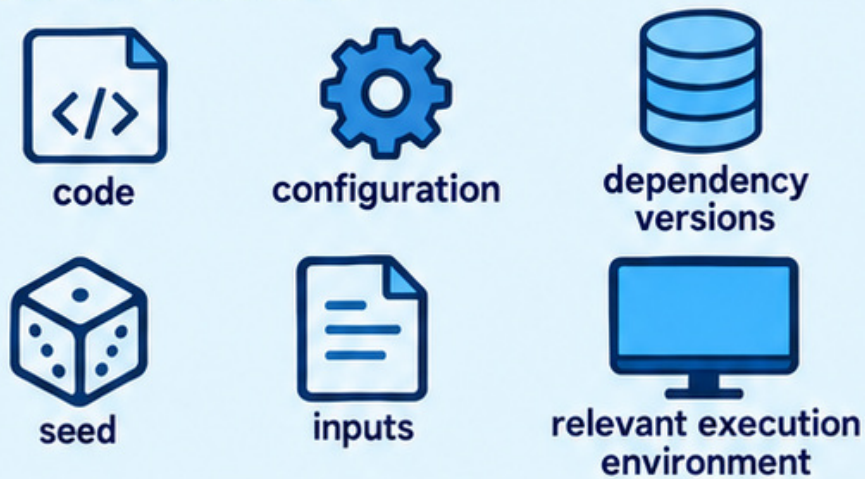


Release Manifest

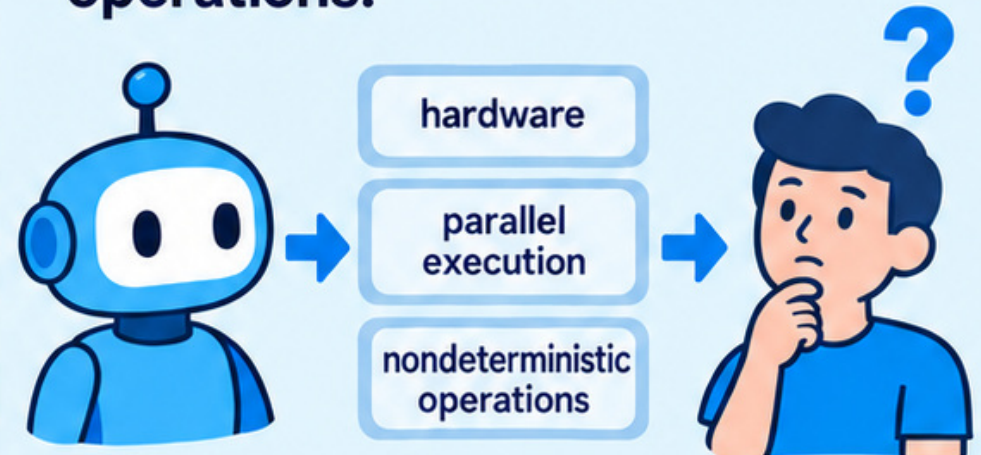
model: Inactivity-model
version: v3
artifact:

REPRODUCIBILITY HAS LIMITS YOU MUST STATE

Record code, configuration, dependency versions, seed, inputs and relevant execution environment.



Repeating a run can still differ across hardware, parallel execution or nondeterministic operations.

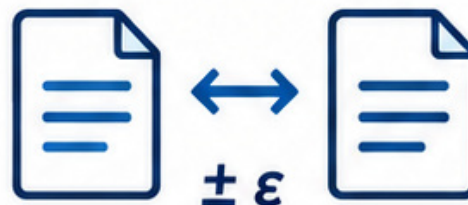


Define the comparison: identical bytes, matching predictions within tolerance, or equivalent measured quality.

Identical bytes



Matching predictions within tolerance



Equivalent measured quality



CHECK THE SAVED ARTIFACT BEFORE ACTIVATION

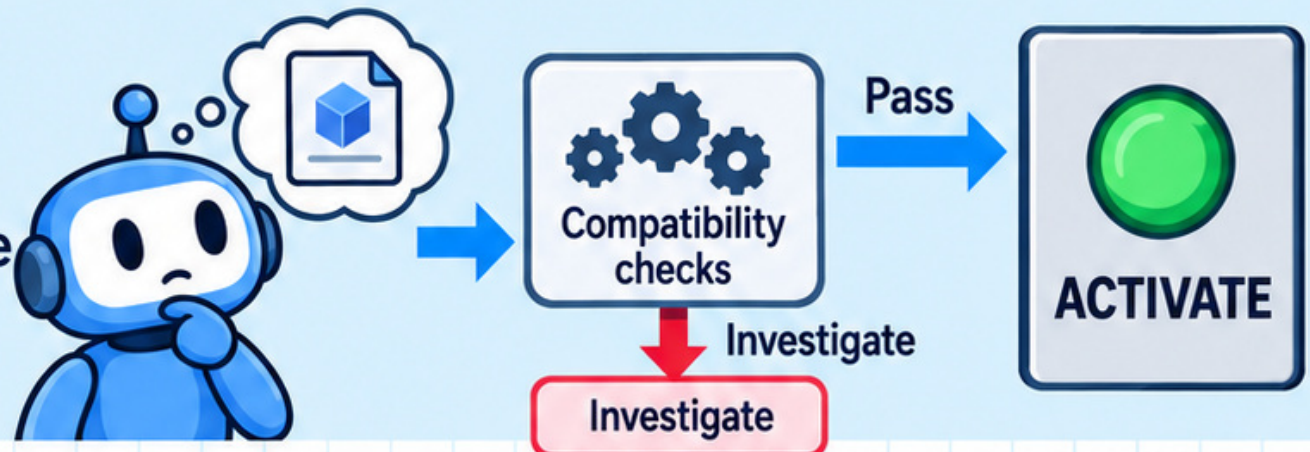
Load the retained artifact in the serving environment.



Run fixed input fixtures through the full pipeline and decision policy. Check outputs, types and missing-feature behavior.

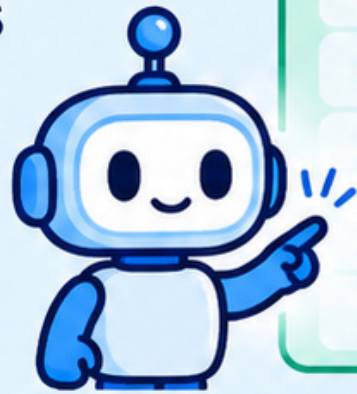


A successful training process does not prove that the packaged release can serve a request.



ROLLBACK THE COMPATIBLE RELEASE BUNDLE

Keep the previous accepted release and its manifest.



Previous (Accepted)

- Model
- Fitted preprocessing
- Feature contract
- Decision policy
- Manifest

Current (Newer)

- Model
- Fitted preprocessing
- Feature contract
- Decision policy
- Manifest

Rollback may need the model, fitted preprocessing, feature contract and decision policy together.



Model



Fitted preprocessing



Feature contract



Decision policy



Check that required feature versions and data remain available before reactivating it.



Required feature versions available



Required data available

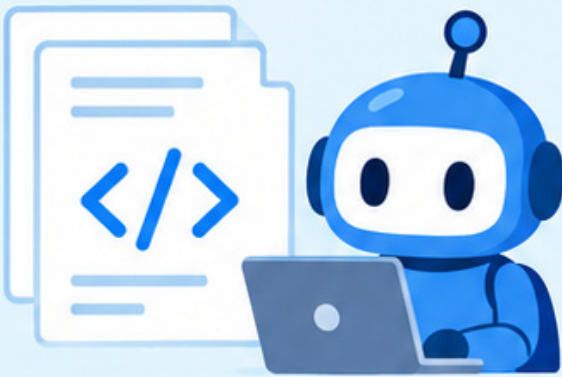


Previous (Accepted)

- Model
- Fitted preprocessing
- Feature contract
- Decision policy
- Manifest

CHOOSE TOOLS FOR DISTINCT RESPONSIBILITIES

Git: source code and reviewed configuration.



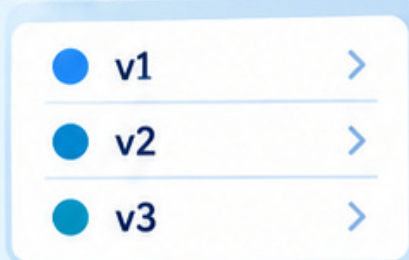
Versioned storage: retained datasets and artifact bytes.



MLflow Tracking: run evidence and comparisons.



Model registry: named versions and references.



Checks in your pipeline: compatibility and acceptance.

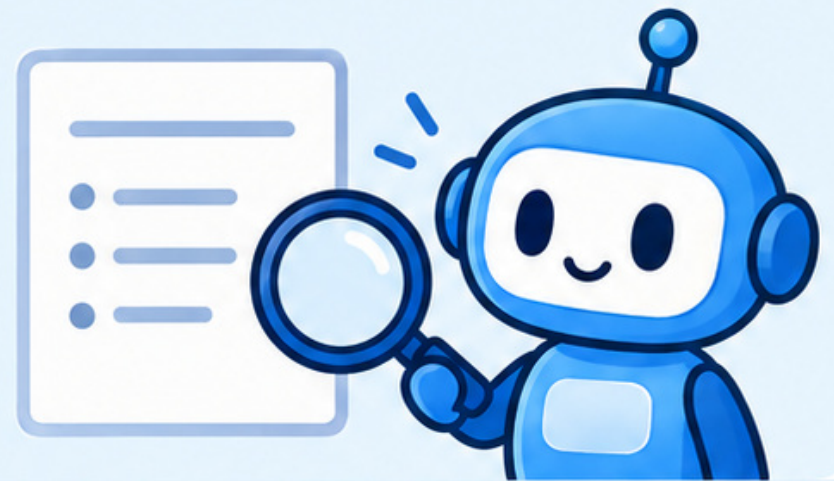


A tool inventory is not a release contract.

YOUR TURN: INVESTIGATE AN INCOMPLETE RELEASE

Toy release record:

- Model filename: model.pkl
- Data source: customers_latest
- Threshold changed from 0.80 to 0.60
- No data revision or decision-policy version



**What behavior
can change?**



**What evidence
is missing?**



**What must be recorded
before a reproducible
release?**

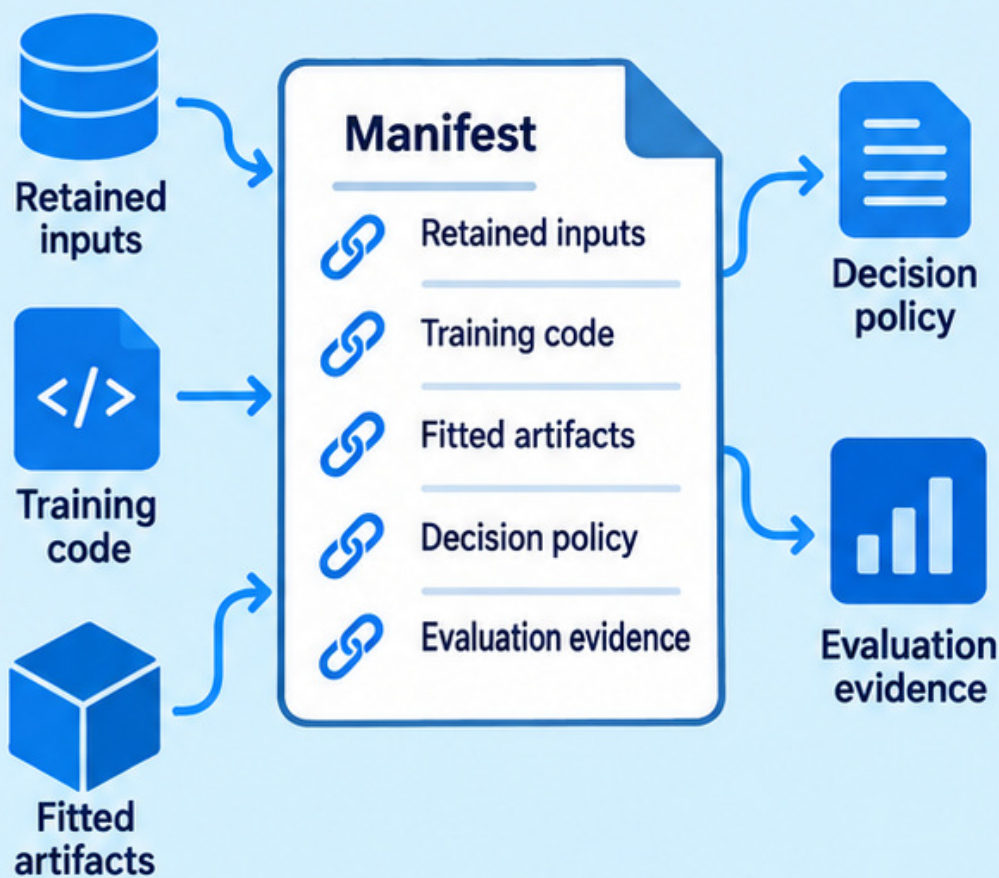


MAKE THE RELEASE INSPECTABLE

A strong release record connects retained inputs, training code, fitted artifacts, decision policy and evaluation evidence.

Start with a small manifest and make every link resolvable.

Release manifest



DAY 13: Batch inference.

Use one pinned release to score a complete cohort reliably.

