



Approve the decision process. Not the demo.

Public evidence frozen · Four judgments sealed · Outcomes added later

Prepared for model risk, markets and treasury leadership, AI governance, and trading technology

What the bank still cannot answer

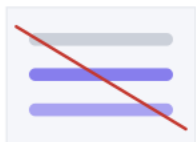
A score can show capability. A demo can show possibility. Neither proves how a specific model behaves on **new market evidence, under declared rules, before the outcome.**

THE GOVERNANCE QUESTION	Static benchmark	Vendor demonstration
Information cutoff	Often implicit	Chosen for the demo
Model, prompt, and tools	Often partial	Shown selectively
Behaviour on new evidence	Not the objective	One curated path
Outcome after the decision	Usually absent	Highlighted later
Record a bank can replay	Task + score	Slides + screenshots

The governance question is prospective: **what did this model see, what did it claim, and can we reconstruct that decision after reality answers?**

The reconstruction problem

Without a fixed cutoff, bound configuration, and sealed output, **a model judgment cannot be reconstructed** by the desk, second line, or a supervisor.



The cutoff was never fixed

The page and surrounding public record have already changed. Nothing proves which information was available when the model answered.



The configuration was partial

The model, prompt, source rules, and tool boundary lived across a demo, a notebook, or someone's memory, not in one run record.



The answer was never sealed

The run produced an output, not a prospective record. Nothing proves the claim existed before the later market outcome was known.

QUESTIONS THAT THEN HAVE NO ANSWER

What was knowable then?

Which configuration ran?

What was claimed before
result?

How should the miss be judged?

Before a model enters a market workflow, a bank should see it make consequential judgments prospectively, under explicit controls, with every answer preserved.

One prospective record for every model

Whalebench gives models the same frozen public item, asks the same four questions, and **seals every answer before the declared outcome is knowable**.

01 Frozen public item

Every model receives the same captured source and the same information cutoff. No later edits or future prices.

02 Four structured judgments

Each model decides relevance, novelty, exposure, and a direction, horizon, and confidence for its market view.

03 Sealed before outcome

The full answer, citations, model, and prompt are preserved while the declared result is still unknown.

04 Two scoreboards

Reasoning quality and market outcome are judged separately, after the forecast horizon closes.



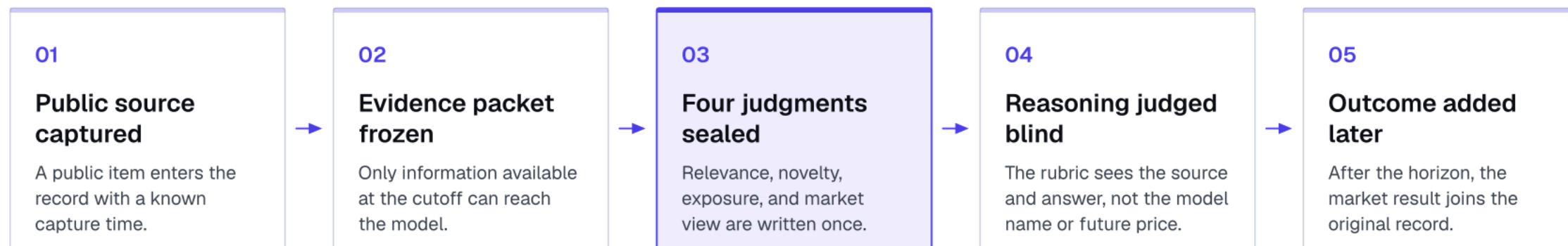
Decision record

Evidence captured	08:12:04 UTC
Information cutoff	08:12:04 UTC
Model + prompt	Versioned
Tools + future data	Disabled
Judgment sealed	08:19:41 UTC
Outcome attached	After the horizon

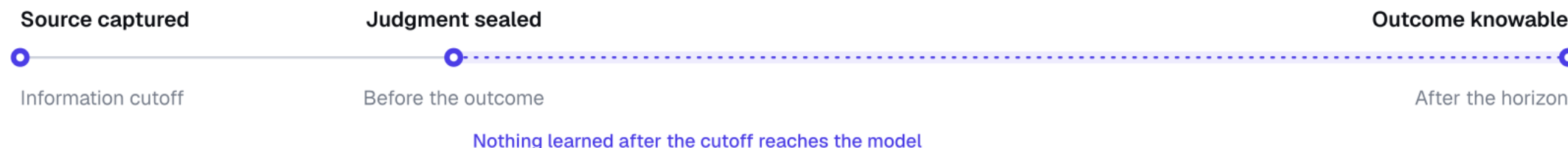
ILLUSTRATIVE RECORD · SHAPE, NOT RESULTS

How a prospective evaluation works

Five steps preserve the distance between what was knowable, what the model claimed, and what happened later.

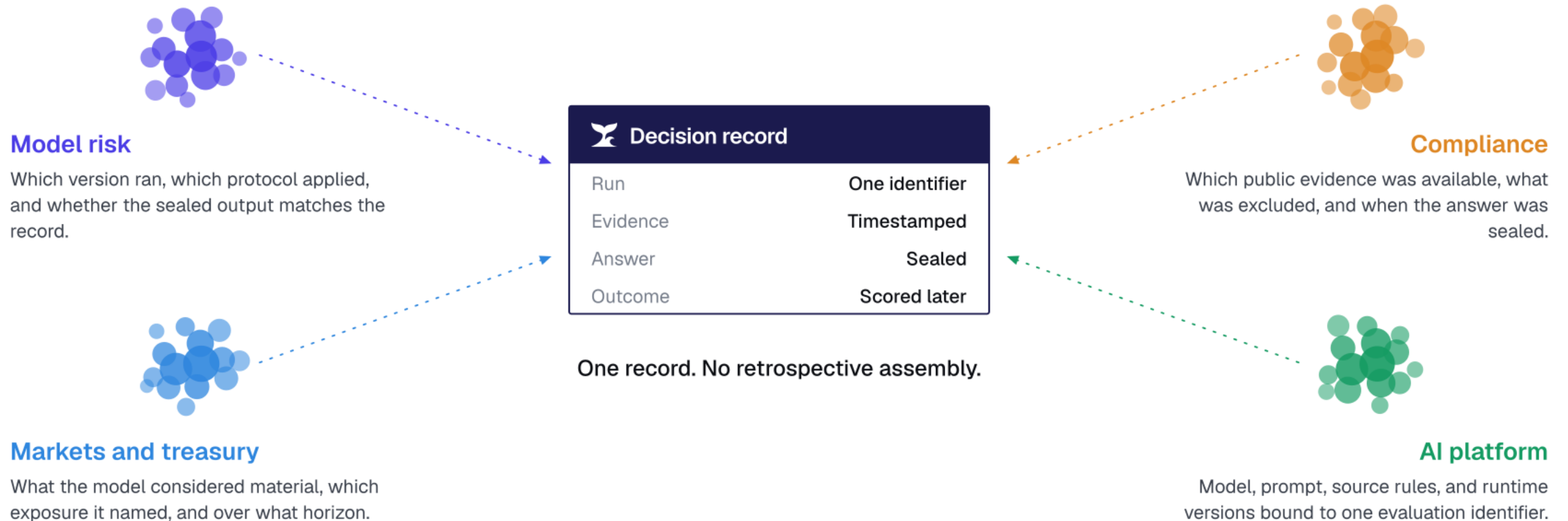


THE SAME RUN ON A CLOCK



One record, four control views

Model risk, markets, compliance, and the AI platform inspect **the same prospective record** instead of reconciling four retrospective reports.



What one prospective run produces

Models answer the same frozen item independently. Every row preserves the configuration, cutoff, and sealed output for later review.

MODEL		VERSIONS IN FORCE	SEALED	STATE
	Claude	prompt v4 · tools v2	08:19:41 UTC	● Sealed
	GPT	prompt v4 · tools v2	08:19:44 UTC	● Sealed
	Gemini	prompt v4 · tools v2	08:19:47 UTC	● Sealed
	Llama	prompt v4 · tools v2	08:19:52 UTC	● Sealed
	Mistral	prompt v4 · tools v1	08:19:58 UTC	● Sealed
	Grok	prompt v4 · tools v2	08:20:03 UTC	● Sealed

Illustrative run record. Timings show the shape of a prospective evaluation, not model results. Outcomes do not exist until the horizon closes.

Prospective

Every output is sealed before its declared market outcome is available.

Versioned

Source, model, prompt, and protocol versions are bound to one identifier.

Inspectable

A review opens the exact captured input and the exact sealed answer.

Exportable

The record can leave as evidence for the bank's own model inventory.

PUBLIC STUDY · REPORTING POLICY

The public pilot records source capture and unranked model judgments. Model-level results publish only after the corpus, rubric, and evaluation window are documented.

The team

Research, product, and market infrastructure stay in one loop. The people running the study are the people a bank works with.



Mark



Krishna



Ansh

Prospective

Every judgment is sealed while its declared market outcome is still unknown.

Unranked today

Model-level results wait for a documented corpus, rubric, and evaluation window.

No future data

Nothing dated after the information cutoff reaches a model before it answers.



Start with one market judgment worth governing.

Public information only. No orders, client data, or future prices.

01 Pick one judgment

Choose a market question your team already reviews today.

02 Declare the protocol

Name the sources, cutoff, output schema, and horizon.

03 Run it prospectively

Seal every answer, then let evidence and the market judge it.

founders@whalebench.com

institutional.whalebench.com