THE FRAMEWORK · PUBLISHED IN FULL

INDEPENDENT MODEL *Review*

A framework for validating credit and decisioning models

PEDRAM NOURANI, PHD, CPA, CA ANZ
VERSION 1.0, AUGUST 2026 PEDRAMNOURANI.COM

WHY THIS FRAMEWORK EXISTS

Models now make decisions that people used to make. Who gets credit, and at what price. Which applications get flagged. Which claims get paid without a human ever reading them. Which transactions get blocked.

Building a model has never been cheaper. Machine learning libraries, vendor decisioning platforms and generative AI have collapsed the cost of production. The cost of verification has not moved. Someone still has to establish whether the model works, whether it is being used the way it was designed to be used, and whether anyone would notice if it quietly stopped working. In a large bank, that job belongs to a model risk team. In most other financial services firms, it belongs to nobody.

Regulators have noticed, and they are not moving in one direction. In April 2026 the US banking agencies replaced SR 11-7, the guidance that defined model risk management for fifteen years, with a lighter and more risk-based framework, and placed generative and agentic AI explicitly outside its scope. Two weeks later, APRA wrote to every bank, insurer and superannuation trustee it regulates and said something close to the opposite: governance is lagging adoption, point-in-time assurance is not keeping up with systems that drift, and risk and internal audit functions must be capable of independently reviewing AI. ASIC has told licensees the same thing since its review of AI use: governance is trailing deployment.

If your firm is not APRA-regulated, none of that binds you directly. It reaches you anyway. Warehouse funders are APRA-regulated banks, and APRA now expects them to see through their supply chains to third and fourth parties. Auditors ask who validated the provisioning model. Boards ask who checked the pricing engine, because directors' duties do not have a technology carve-out. The question always arrives from outside, and it is always the same question: who reviewed the model, and what did they find?

This framework is my answer to how that review should be done. It is published in full because the methodology is not the secret. Any firm can run these questions internally, and firms that do will be better for it. What an independent review adds is not the checklist. It is the independence.

WHAT COUNTS AS A MODEL

A model is any method that turns input data into an output someone relies on for a decision. That includes credit scorecards, application and behavioural scoring, pricing engines, provisioning and expected loss models, serviceability calculators, fraud and transaction monitoring systems with learned components, and generative AI wherever its output enters a decision.

Two boundary cases matter more than the definition.

Vendor models are in scope. You can outsource the build. You cannot outsource the accountability. If a vendor's model declines an applicant, the licence obligations, the funder covenants and the customer outcome all still sit with you. A vendor's marketing document is not validation evidence.

Tools that "assist" are in scope once anyone stops checking. A model sold as decision support becomes the decision maker on the day the human review becomes a formality. A review should test what actually happens, not what the policy says happens.

TIER BEFORE YOU TEST

Not every model deserves the same depth of review, and pretending otherwise is how validation becomes a box-ticking cost that everyone resents. Two dimensions set the tier: materiality, meaning how much exposure the model influences, and autonomy, meaning whether the model decides or merely advises.

TIER 1

The model decides, and the exposure is material. Automated credit decisioning, pricing, provisioning. Full review, annually or on material change.

TIER 2

The model advises on material exposure, or decides on small exposure. Full review on a longer cycle, with monitoring checked in between.

TIER 3

Low materiality and low autonomy. An inventory entry, a named owner, basic controls. Nothing more.

The first artefact of any engagement is the model inventory with tiering, because most firms are running more models than they think they are. Finding the ones nobody listed is usually where the review starts earning its fee.

THE FIVE QUESTIONS

Every review under this framework answers five questions. For each one, this section sets out what good looks like, what a review tests, and where models usually fail.

QUESTION 01

Is the model designed for the job it is doing?

WHAT GOOD LOOKS LIKE

A documented objective. A defined target variable and a defined population. Stated assumptions and stated limitations. Development evidence that was retained, not reconstructed. Some record that alternatives were considered.

WHAT A REVIEW TESTS

Whether the documentation describes the model that exists. Whether the target variable and development population still match the decision the model is making today. Whether the methodology fits the problem, or was simply the method the developer knew.

WHERE IT USUALLY FAILS

The model was built for one purpose and drifted into another. A target variable defined on a 2021 portfolio is still scoring a 2026 through-the-door population that looks nothing like it. The documentation was written after the model went live, to describe what was already there.

QUESTION 02

Is the data fit to carry it?

WHAT GOOD LOOKS LIKE

Lineage from source system to model input that someone can walk you through. Quality controls at the point of entry. Evidence that the development sample represents the population the model now scores. Leakage checks. For third-party data, stable definitions and clear usage rights.

WHAT A REVIEW TESTS

The lineage, walked end to end. A data quality profile on the current inputs. A comparison of the development population against the current population. Whether any input would not be available, in that form, at the moment of a real decision.

WHERE IT USUALLY FAILS

The development sample comes from a benign period and the model has never seen a downturn. An upstream system changed a field's meaning and nobody told the model owner. A variable performs suspiciously well because it leaks the outcome.

QUESTION 03

Does it perform, and how would you know?

WHAT GOOD LOOKS LIKE

Discrimination, calibration and stability measured on a schedule, against thresholds that a named person owns. Backtesting against realised outcomes. Out-of-time testing, not just out-of-sample. Benchmarking against a challenger or a simple alternative. Override analysis.

WHAT A REVIEW TESTS

The headline metrics, recomputed independently from raw data rather than read from the monitoring pack. Population stability. Calibration against what actually happened. Override rates, and how overridden decisions performed.

WHERE IT USUALLY FAILS

The monitoring pack has reported the same three green metrics for three years while the portfolio underneath it changed. Stability indicators sit at amber with no owner and no trigger for action. Overrides outperform the model, which means staff have privately stopped trusting it, or underperform it, which means the exception process is the real risk.

QUESTION 04

Is the model in production the model that was approved?

WHAT GOOD LOOKS LIKE

Production configuration reconciled against approved documentation. Controls on inputs. Version history and a change log that is actually maintained. Monitoring wired to the production system, not to a copy of it.

WHAT A REVIEW TESTS

A reconciliation of what runs against what was approved. A sample of real decisions traced end to end, from input to outcome. The change management record, including vendor-initiated changes.

WHERE IT USUALLY FAILS

Cut-offs were adjusted in production during a busy quarter and never written down. The vendor pushed two model updates last year and nobody at the firm can say what changed. There is a spreadsheet stage in the middle of the pipeline that no diagram mentions.

QUESTION 05

Who owns it, and what happens when it breaks?

WHAT GOOD LOOKS LIKE

A named owner with the authority to act. A review cadence set by tier. A defined escalation path. A fallback that has actually been tested. Reporting that reaches the board or the funder in a form they can challenge.

WHAT A REVIEW TESTS

The governance in practice, through interviews as much as documents. Whether escalation has ever fired, and what happened when it did. Whether the fallback is credible at production volume. Whether upward reporting communicates or conceals.

WHERE IT USUALLY FAILS

Ownership sits with a committee, which means it sits with nobody. The person who built the model left, and the knowledge left with them. Board reporting summarises so aggressively that the problem is invisible. The fallback is "we would process manually", for a volume no team could process manually.

A NOTE ON GENERATIVE AI

Generative systems do not fail the way scorecards fail, but they answer the same five questions. What changes is the evidence. You cannot compute a Gini coefficient on a chatbot. In its place: a precisely defined task, a ground-truth test set built for that task, an error taxonomy that records what kinds of wrong the system produces and at what rates, version control over models and prompts, and testing of whether human oversight actually catches errors at production volume rather than in a demo.

If a generative system's output enters a credit decision, a hardship response or a disclosure document, it is a Tier 1 or Tier 2 model and should be reviewed as one. The US carve-out of generative AI from model risk guidance means there is no rulebook for it. It does not mean there is no risk in it.

WHAT AN INDEPENDENT REVIEW INVOLVES

A review is fixed in scope and fixed in price: one model or decisioning system, three to four weeks, delivered remotely. It runs in three stages.

Stage 1: scope and evidence, week 1. Confirm the model under review and its tier. Agree the evidence list and access. Schedule interviews with the model owner, the developer or vendor contact, and the people who use the output. A typical evidence request:

- model documentation and development evidence
- development data and recent performance data, or extracts sufficient to recompute headline metrics
- monitoring packs for the past twelve months
- the change log and version history
- override and exception reports
- vendor documentation and contract terms for third-party models
- relevant policies covering model risk, data and AI use

Stage 2: testing, weeks 2 to 3. The five questions, run against the evidence. Independent recomputation of performance wherever the data allows it. Interviews to test whether governance on paper is governance in practice.

Stage 3: reporting, week 4. Draft findings issued for management response. Final report and attestation delivered.

Independence. Findings are addressed to the board, the audit committee or the funder, not to the model owner. I do not remediate findings I have raised within the same engagement, for the same reason an auditor does not audit their own bookkeeping. Client data stays in the client's environment wherever practicable. Where extracts are required, they are minimised, agreed in writing, and destroyed on completion.

WHAT YOU RECEIVE

Three documents.

The validation report. Written for a board, not for a data scientist. Each finding stated plainly, with the evidence behind it and what it means for reliance on the model.

The findings register. Every finding rated by severity:

— **HIGH**

The finding undermines reliance on the model's output, or creates a breach of an obligation. Address before continued reliance.

— **MEDIUM**

A weakness that degrades performance or control but does not yet undermine reliance. Address within an agreed window.

— **LOW**

An improvement opportunity.

The attestation of scope. One page. What was reviewed, as at what date, on what evidence, and what was not reviewed. This is the page funders and auditors actually ask for, because it is the page they can rely on and file.

WHAT THIS FRAMEWORK IS NOT

A review under this framework is a point-in-time opinion on a defined scope. It is not remediation, not model development, not a guarantee against loss, and not legal advice. A model that passes today can fail next year. That is exactly why question five exists: a point-in-time review of a system that drifts is only as good as the monitoring it verifies. Tier 1 models should be re-reviewed annually or on material change, whichever comes first.

USING THIS FRAMEWORK

Use it. Run the five questions against your most material model this month. If the answers hold, write them down, because that record is worth having before a funder, an auditor or a director asks for it. If the answers do not hold, you have found the work.

If you want the questions answered independently, that is the service. Contact details are at pedramnourani.com.

REGULATORY REFERENCES

1. Board of Governors of the Federal Reserve System, OCC and FDIC, Revised Guidance on Model Risk Management (SR 26-2, OCC Bulletin 2026-13, FIL-15-2026), 17 April 2026, superseding SR 11-7 (2011).
2. APRA, Letter to industry on Artificial Intelligence, 30 April 2026.
3. APRA, Prudential Standard CPS 230 Operational Risk Management, in force from 1 July 2025.

ABOUT THE AUTHOR

Pedram Nourani, PhD, CPA, CA ANZ

Pedram Nourani holds a PhD in Economics and the CPA (Australia) and CA ANZ designations. His peer-reviewed research covers bank efficiency, risk and financial technology. He writes Capital & Code, a newsletter on how AI is changing the work of finance. Contact: info@pedramnourani.com.