

ALPHA IQ · CERTIFY IQ · WHITE PAPER

Model Checks You Can Prove

Evidence-traced, repeatable validation of the models banks and insurers rely on

Ilan Gleiser¹

¹AlphaIQ — ilan@alphaiq.dev

8 October 2026

Abstract

Banks and insurers run on models: who gets a loan, what a policy costs, how much capital to hold, which transactions look suspicious. Regulators require every one of those models to be checked, and the check itself to be defensible. Today that work is slow, manual and hard to reproduce: a validator rebuilds spreadsheets, re-types numbers into a report, and the result cannot be traced back to the data it came from. CertifyIQ is a platform that performs these checks from evidence. It fingerprints every file on arrival, runs the tests that apply to the model's family against thresholds that come from a published rule or the institution's own written policy, writes the report in the format the supervisor expects, and only issues a certificate once a named owner has signed for the evidence. Every figure in the output is traced to the inputs it was computed from; anything that could not be established is stated rather than omitted. We describe the design, walk through a complete validation of an IRB rating system under the ECB TRIM template, and set out what the platform covers today and what it does not.

Keywords: *model risk management, model validation, IRB, ECB TRIM, SR 26-2, EU AI Act, evidence provenance, reproducibility, certification*

1 Why this matters

A bank decides who gets a loan with a model. An insurer prices a policy with a model. A trading desk decides how much capital to hold with a model. A compliance team flags suspicious payments with a model. When one of those models is wrong, people are turned down who should not be, prices are set too low to cover the claims that arrive, capital runs short, or crime goes unseen. Because the stakes are this high, regulators require every such model to be checked, and they require the check itself to be defensible: a supervisor must be able to see what was tested, against what, by whom, and on which data [1, 2].

In practice the check is slow and fragile. A validator receives spreadsheets and documents, rebuilds pieces of the model to see whether it behaves as described, runs statistical tests by hand, and types the numbers into a report. It takes weeks. Nobody can be entirely sure the

numbers in the report are the numbers that came out of the tests, because the two are joined only by a person's attention. When the regulator asks a year later how a figure was arrived at, the answer is often an archaeology exercise.

The requirement is also widening. Machine-learning models, generative-AI systems and whole decision processes are now in scope: the EU AI Act's technical documentation duties for high-risk systems [3], the NIST AI Risk Management Framework [4], the ECB's guide to internal models [5], and the Federal Reserve's model risk guidance [1] all ask for evidence, not assurance. A validation function that cannot show its work is a validation function that cannot keep up.

CertifyIQ was built on one premise: a model check should be something you can prove. Not a PDF that asserts a model is fine, but a chain from the raw files, through the tests that ran, to the sentence in the report, with every link recorded and any link that is missing named as missing.

2 What CertifyIQ does

CertifyIQ is a platform for validating the models that banks and insurers rely on. It does four things, in order, and each one leaves evidence that the next one rests on.

It collects the evidence. Model documents, data extracts, outcome files and the owner's own declared thresholds are uploaded once and given a cryptographic fingerprint on arrival. The fingerprint is the file's content hash: change one row and the hash changes, and the platform can tell every report that depended on the old version.

It imports and cleans the data in stages. A file is checked for structure, parsed, mapped column by column with a named person approving the mapping, screened for personal data, normalised to a canonical form and fingerprinted again as the version the tests will run on. What was rejected, and why, is recorded as evidence alongside what was kept.

It runs the checks that apply to the model's family. A credit scorecard, an IRB rating system, a fraud model, an AML alert engine, a market-risk VaR model, a machine-learning classifier, an insurance severity model, an interest-rate-risk model, a funding and liquidity model and a generative-AI system each have their own pack of tests. Each test is judged against a threshold that comes from a published rule or from the institution's own written policy. Where neither exists, the figure is reported and the row is marked as not judged.

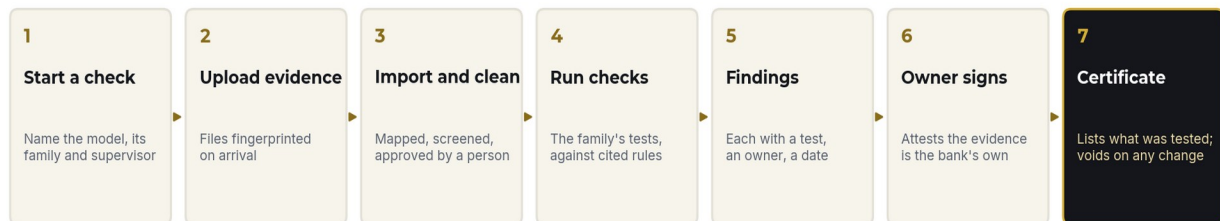
It writes the report and tracks the findings. The same evidence produces the report in the format the supervisor expects (Table 1), each finding is given an owner, a severity and a due date, and the certificate that follows lists exactly what was tested and what was not.

Table 1. Report formats written from one body of evidence. Every figure in every format carries the hash of the data it was computed from.

Format	Written for
SR 26-2	US banks: the Federal Reserve's model risk management expectations
ECB TRIM / IRB review pack	Euro-area banks: the ECB's internal-model guidance for IRB rating systems
EU AI Act Annex IV	Technical documentation for a high-risk AI system
NIST AI RMF	A risk-management assessment of an AI system
FRTB	Market-risk internal models: back-testing and P&L attribution
ASOP 56	Actuarial model governance for insurers
Two-page summary	A board or committee that needs the verdict, the findings and what is outstanding

3 How it works

A validation is seven steps (Figure 1). The order matters, because each step produces an artefact the next one is computed from, and the certificate at the end names the hash of each artefact it rests on.



Every artefact produced along the way is content-addressed; the certificate names the hash of each one it rests on.

Figure 1. The seven steps of a CertifyIQ validation. Each step is content-addressed; the certificate at the end is only issued on evidence a named owner has signed for.

3.1 Evidence is content-addressed

Every artefact — a file, a parsed population, a column mapping, a declared threshold, a test run — is stored under the hash of its canonical serialisation. Keys are sorted, numbers are written in their shortest round-trip form, and absent values are dropped, so two artefacts that mean the same thing serialise to the same bytes. A test run's hash covers everything that can move its result: the data it read, the model specification, the parameters and the version of the code that ran. It deliberately does not cover the wall clock or a row identifier, so the same evidence really does hash the same.

$$h(A) = \text{SHA256}(\text{canon}(A)), \quad h(\text{run}) = \text{SHA256}(h(D), h(M), h(\theta), h(\text{code})) \quad (1)$$

The consequence is that a run can be classified against what is already stored before anybody decides to record it. A run that reproduces a stored one is recorded as reproducing it, not as a new result. A run whose figures are identical but whose verdicts moved, because a threshold was declared in between, is shown as exactly that. A run that diverges is a divergence, and the report says so.

3.2 Thresholds come from a rule, or from you — never from us

Where a regulator publishes a limit, the test uses it and the report cites it: the PD floor in Article 160 of the Capital Requirements Regulation [6], the Basel back-testing zones for VaR exceptions. Where no limit is published, the institution's own documented threshold is used, with the document and the person who set it recorded next to the figure. Where there is no rule and no policy, the figure is shown without a pass or fail. CertifyIQ never invents a threshold to make a table look complete, and the IRB example in Section 4 shows how often that case arises in practice.

3.3 A person signs for the evidence

A certificate asserts that named tests were performed on named evidence. Evidence nobody has claimed is not named evidence: it is a file the platform was handed. So before a certificate is issued, the owner of each input artefact at the institution attests that it is theirs, and the attestation names the hash it covers. The validator who signs the certificate cannot be the person who attested the evidence. If an artefact that every figure rests on is unattested, no certificate is issued; if an artefact that only some figures rest on is unattested, the certificate is issued with those figures outside its boundary, and it says which ones and why.

3.4 The certificate voids itself

An issued certificate carries the hashes of the model specification, each evidence artefact and the report. If any of them changes, the certificate no longer matches and is void. It also expires. It cannot be edited after issue; a withdrawal is recorded as a new fact, not by deleting the old one. Anyone with the certificate can check later whether it still holds. Four outcomes are possible: certified, certified with conditions, not certified, and incomplete, the last meaning a declared part of the scope has neither been tested nor deliberately excluded.

4 A worked example: an IRB rating system under the ECB TRIM template

To make this concrete we validated a corporate probability-of-default rating system end to end on the platform, using a synthetic evidence pack built for the purpose [7]. The pack is

deliberately imperfect in one known place so that the report has something to find. Nothing in it describes a real bank, borrower or facility.

4.1 The evidence

Five files were uploaded. A master rating scale of seven grades, AAA-AA to CC-C, each with the lower and upper one-year default probability the bank assigns to it, declared from page 31 of the model development document by a named person (Figure 2). A graded population of 4,000 obligors with grade, score and a default flag, 66 of them defaulters, observed as at 30 June 2026. The bank's published one-year transition matrix (table 4.7 of the model document) and a migration book of 11,140 obligors observed at both 30 June 2025 and 30 June 2026. And a facility book of 2,000 facilities with the loss-given-default the model estimated and the loss that was realised.

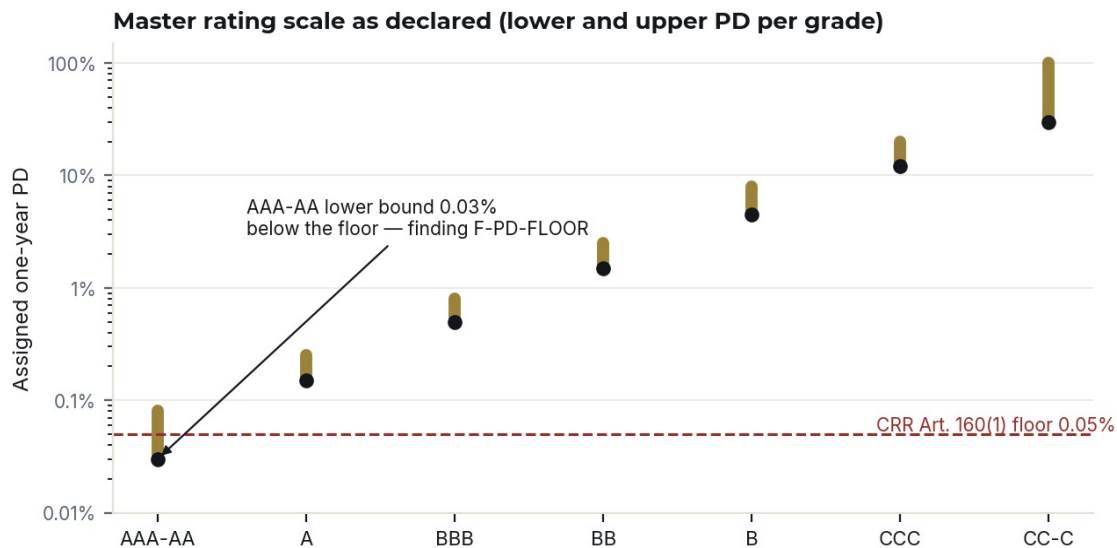


Figure 2. The declared master rating scale. The lower bound of the best grade sits below the regulatory floor; this is the one planted defect in the evidence pack, and the one material finding the report raised.

4.2 What the report found

The engagement was registered as an IRB rating system under the EU supervisory frame, which routes the memo to the ECB TRIM / IRB review pack template. Ten tests ran. Table 2 shows the results as the report printed them.

Table 2. Results of the IRB validation as printed on the TRIM-template report. Rows marked 'reported' carry no verdict because neither the ECB nor the engagement declared a threshold.

Test	Result	Verdict	Judged against
PD floor (per grade)	AAA-AA assigned 0.030%	Material finding	CRR Art. 160(1): 0.050%
Jeffreys calibration back-test	weakest grade p = 0.037	reported	no ECB significance level

Test	Result	Verdict	Judged against
Predicted vs observed default rate	largest deviation 207 bp	reported	no threshold declared
Herfindahl grade concentration	H = 0.512	reported	no threshold declared
Rating migration stability	weakest cell AAA-AA→BBB, p = 0.0575; 31 cells testable, 33 untested	reported	no threshold declared
LGD predictive ability (paired t)	mean diff -0.0014, t = -0.652, p = 0.514	reported	no threshold declared
LGD discriminatory power (gAUC)	0.856	reported	no threshold declared

The one material finding, F-PD-FLOOR, is the planted defect: the best grade is assigned a probability of default of 0.030 per cent, below the 0.05 per cent floor that Article 160(1) of the CRR sets for corporate exposures [6]. The report cites the article, names the grade and the figure, and the finding enters the register with an owner to be assigned.

The remaining rows illustrate the threshold rule in Section 3.2. The ECB's internal-model guidance describes the Jeffreys test, the Herfindahl index and the AUC-stability comparison, but publishes no pass/fail level for them [5], and this engagement declared none of its own. CertifyIQ therefore printed the figures and marked the rows as reported, not judged. Nine of the ten rows on an IRB report fall into this class until the institution declares thresholds under the engagement's claims, at which point the same stored figures are re-judged without re-running anything. We regard this as the correct behaviour: a validation platform that supplied its own significance levels would be making a supervisory judgement that is not its to make.

4.3 The tests, in plain terms

The Jeffreys back-test asks, for each grade, how surprising the observed defaults would be if the assigned PD were right. With d defaults among n obligors in grade g , the posterior for the true default rate under a Jeffreys prior is a Beta distribution, and the reported p-value is the probability that the true rate is at or below the assigned one [8]:

$$p_g = P(\pi \leq \text{PD}_g \mid d_g, n_g), \quad \pi \sim \text{Beta}(d_g + \frac{1}{2}, n_g - d_g + \frac{1}{2}) \quad (2)$$

Herfindahl concentration measures whether the obligors are spread across the grades or bunched into a few. With share s of the population in each grade, the index is the sum of squared shares; a value of 0.512 means that roughly half the population's weight sits in the

largest grades, which the report prints so that a reader can judge whether the scale is doing its job of discriminating:

$$H = \sum_{g=1}^G s_g^2, \quad s_g = \frac{n_g}{\sum_k n_k} \quad (3)$$

LGD predictive ability is a paired comparison of what the model estimated against what was actually lost on each facility. A mean difference of -0.0014 with $t = -0.652$ and $p = 0.514$ says the estimates are not biased in either direction on this book. The gAUC of 0.856 is a separate question — whether the model ranks facilities by severity of loss — and a value below one says it does so imperfectly. The pack ships two further variants of the books in which the migration of one grade has drifted and every LGD estimate is understated by three points, so a reader can confirm that the stability test flags only the drifted row and the paired t-test catches the bias while the gAUC does not move.

$$t = \frac{\bar{\Delta}}{s_{\Delta}/\sqrt{n}}, \quad \Delta_i = \text{LGD}_i^{\text{est}} - \text{LGD}_i^{\text{real}} \quad (4)$$

4.4 The certificate

The report's recommendation was that the model may be relied upon, with conditions. The certification screen nonetheless refused to issue a certificate, and the refusal is the most instructive part of the exercise (Figure 3). Nobody at the institution had attested any of the four input artefacts. The graded population and the master rating scale are what every figure rests on, so their absence blocks the certificate outright; the migration and facility books carry three metrics between them, so their absence would only narrow it. The screen says exactly which is which, names the metrics that would fall outside the boundary, and asks for the owners to attest before issue.

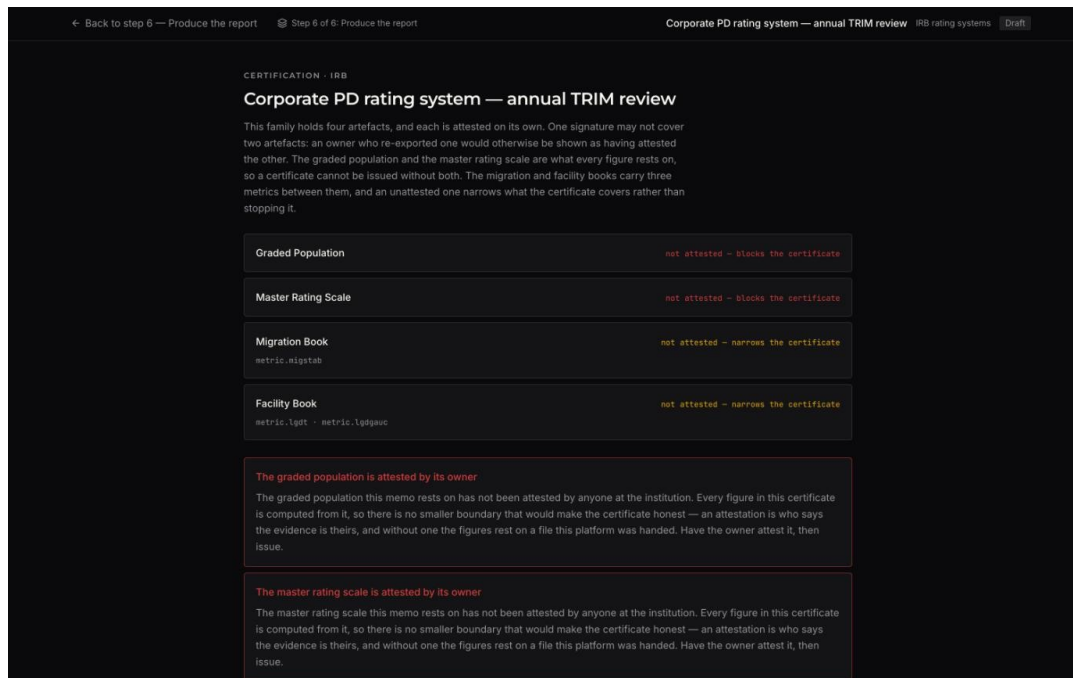


Figure 3. The certification screen for the IRB engagement. Two unattested artefacts block the certificate, two narrow it, and the omitted metrics are named. No certificate is issued on evidence nobody has claimed.

This is the design working as intended, and it is also where a gap showed: the IRB attestation can be seen on this screen but not yet performed from it, so the four signatures have to be collected through the generic attestation path until the dedicated flow ships. We record it here rather than present a certificate the platform did not issue.

5 What makes it different

It shows its work. Every sentence about a figure says where the figure came from, what data version, what threshold and whose declaration. A supervisor can follow any number on the report back to the bytes it was computed from.

It refuses to guess. Missing data is reported as missing. An unpublished threshold is reported as unpublished. A test that could not run is named, with the reason. A machine-learning model is never described as a replica of anything; a rebuild starts only from what the institution wrote down. If the documents do not support a claim, the report does not make it.

It is the same every time. Same inputs, same version, same numbers. Tests and scenarios run on pinned inputs with a fixed random seed, and any stored result can be re-run and checked against its hash.

A human stays in charge. A named person approves the column mapping, declares each threshold, attests each artefact and signs the certificate. The platform computes; people decide, and the record says which people.

6 Benefits

A validation that takes weeks of a validator's time becomes a matter of hours once the evidence is on file, because the tests, the thresholds and the report are produced rather than assembled. It is cheaper for the same reason. It is provable, because every figure is traced to its inputs and the certificate voids itself if those inputs change. It is consistent, because the same evidence produces the same report, and the report for a second regulator is written from the same figures rather than re-derived. And it is ready for new rules: when a supervisor publishes a threshold that was previously missing, declaring it re-judges every stored figure without re-running a single test.

7 Who uses it

Model owners upload the evidence for their model, declare the thresholds their policy sets, and attest that the data is theirs. They see the findings raised against their model and the due dates attached to them.

Validators run the checks, read the results, assign severities and owners to findings, and sign the certificate. They cannot attest the evidence they validate.

Approvers read the two-page summary and the findings register and decide whether the model may be relied upon, and on what conditions.

Auditors and supervisors receive a report in the format they expect, with the run reference printed on it, and can check any figure against the evidence it names, and any certificate against the hashes it carries.

8 What it covers today

Ten model families have working checks on the platform (Table 3). Each family has its own intake, its own pack of tests and its own report template, and a sample report for each is published without sign-in so that a reader can judge the output before uploading anything. Two further families are on the roadmap and are not built; we list them so that the boundary is stated rather than implied.

Table 3. Families with working checks, and the two that are not yet built. GenAI systems are judged by a second model whose agreement with human labels must reach a Cohen's kappa of at least 0.70 before its scores may be reported.

Family	What the checks measure	Status
Credit scorecards	Discrimination, calibration, stability, severe-scenario PD, margin of conservatism	Working
IRB rating systems	PD floor, Jeffreys back-test, concentration, migration stability, LGD bias and ranking	Working

Family	What the checks measure	Status
Fraud models	PR-AUC, recall at review rate, feature leakage, drift, disparate impact	Working
AML alert engines	Detection rate, alert-to-SAR conversion, below-the-line testing, typology coverage	Working
Market risk (VaR / ES)	Back-testing exceptions, Kupiec and Christoffersen tests, ES error, P&L attribution	Working
Machine-learning classifiers	Performance, stability, fairness and documentation checks on the owner's own scores	Working
Insurance severity	Fit and tail behaviour of claim-severity models	Working
Interest-rate risk (IRRBB)	EVE and NII sensitivities against the supervisory outlier tests	Working
Funding and liquidity	Liquidity and funding metrics for ICAAP / ILAAP reporting	Working
Generative-AI systems	Claim support, citation precision, tool-call scope, contradictions, via a proven judge	Working
Actuarial reserving	—	Roadmap, not built
Simulation models	—	Roadmap, not built

Beyond single models, an engagement can declare the whole decision process a model sits in — data, models, rules, human review and the action the customer receives — with each link needing evidence and each hand-off an owner. Until every link is tested or deliberately excluded, the certificate for the process says incomplete.

9 Limits

Three limits are worth stating plainly. First, a report is only as good as the evidence declared: CertifyIQ tests what it is given and names what it was not, but it cannot discover a data source nobody uploaded. Second, for several IRB tests there is no published threshold, so the verdict column stays empty until the institution declares one; the platform makes this visible rather than filling it in, and some readers will find a report with unjudged rows uncomfortable. Third, the platform runs today as a managed, multi-tenant cloud service with database-level isolation per client and an automated check that proves the separation; deployment inside an institution's own environment is open and decided case by case, not a shipped option.

10 Next step

Start with one model that is due for review. Upload its evidence, let the platform run the checks, and compare the report with the last validation you filed. The evidence and findings stay yours

and can be exported at any time. Sample reports for every family are open at certifyiq.dev, and the IRB evidence pack used in Section 4 is available on request so the walk-through can be repeated.

CERTIFYIQ BY ALPHAIQ — INDEPENDENT MODEL VALIDATION

Every figure is traced to the evidence it was computed from, and what was not established is stated rather than omitted.

References

- [1] Board of Governors of the Federal Reserve System. Supervisory Letter SR 26-2, Guidance on Model Risk Management. Washington, DC, 2026.
- [2] Basel Committee on Banking Supervision. Basel III: Finalising post-crisis reforms. Bank for International Settlements, December 2017.
- [3] European Parliament and Council. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Annex IV: Technical documentation. Official Journal of the European Union, 2024.
- [4] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1, January 2023.
- [5] European Central Bank. ECB Guide to internal models. Frankfurt am Main, consolidated edition, 2024.
- [6] European Parliament and Council. Regulation (EU) No 575/2013 on prudential requirements for credit institutions (Capital Requirements Regulation), Article 160(1). Official Journal of the European Union, 2013, as amended.
- [7] AlphaIQ. CertifyIQ IRB evidence pack for the ECB TRIM template: synthetic master rating scale, graded population, transition matrix, migration book and facility LGD book. October 2026.
- [8] Jeffreys, H. An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society A, 186(1007):453–461, 1946.
- [9] Hanley, J. A. and McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1):29–36, 1982.
- [10] Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1):37–46, 1960.
- [11] European Banking Authority. Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures (EBA/GL/2017/16). November 2017.
- [12] Actuarial Standards Board. Actuarial Standard of Practice No. 56, Modeling. December 2019.