

Annex IV technical documentation — Application Scorecard Challenger v4

EU AI Act — Annex IV technical file

EU AI Act high-risk system dossier

MODEL

Application Scorecard Challenger v4

ENGAGEMENT

Sample — Annex IV technical file, application scorecard

PREPARED

2026-10-07

PREPARED BY

AlphaIQ CertifyIQ, computed by the product

RECOMMENDATION

● May be relied upon, with conditions

10 sections · 6 traced figures · 2 finding(s) · 19 test(s) named as not performed

SYNTHETIC SAMPLE

Every figure in this report was computed by CertifyIQ from a synthetic book — The same scored book as the ML sample, under a declared EU supervisor — so the same evidence becomes the Annex IV dossier. No real institution, model or customer is described. The book is `ml-scored-population.csv`; how it was generated, and the defects planted in it, are documented in `ml-scored-population-dgp.json`.

AT A GLANCE · SYNTHETIC SAMPLE

Application Scorecard Challenger v4

EU AI Act — Annex IV technical file · EU AI Act high-risk system dossier

May be relied upon, with conditions

1 material finding(s) require remediation or an accepted compensating control before unqualified reliance. Each is listed in section 9 with the evidence that produced it. 5 of 6 rows were not judged (no threshold declared) and are outside this conclusion.

10

SECTIONS

6TRACED
FIGURES**1**

MATERIAL

1

NOTABLE

0

OBSERVATIONS

19NOT
PERFORMED

Supervisory frame: this memo is built to "EU AI Act — Technical Documentation", published under EU · Annex IV. The engagement declares supervision by EU, which is the jurisdiction this frame is published under.

Evidence manifest: none sealed. The documents behind each figure are named where the figure is; the set of documents on record when this memo was written has not been attested by anyone, so it cannot be cited here.

How to read this sample. Every number below was computed by CertifyIQ from a synthetic book (mL-scored-population.csv), through the same resolver, metric pack and report builder a signed-in engagement uses. Each figure carries its own trace — where it came from and what it was judged against. Tests that could not be performed are named with the reason, never left out. Where no threshold was declared, a metric is reported but not judged, because this report does not invent an institution's tolerances.

CONTENTS

- | | |
|---|---|
| 1 General description of the AI system | 6 Relevant changes made through the system's lifecycle |
| 2 Development, data governance, and the metrics used to measure accuracy | 7 Harmonised standards applied, or the alternative solutions adopted |
| 3 Monitoring, functioning and control, including accuracy for specific groups | 8 EU declaration of conformity |
| 4 Appropriateness of the performance metrics | 9 The post-market monitoring plan and what it has reported (Article 72) |
| 5 Risk management system (Article 9) | 10 Coverage of this dossier, and what it does not evidence |

1 General description of the AI system

This dossier covers Application Scorecard Challenger, version v4, a machine-learning system operated by AlphaIQ (internal sample). Annex IV(1) asks what the system is for and which version is documented; both are read from the engagement record rather than restated here, so a dossier cannot describe a version other than the one its evidence was drawn from.

The unit of observation is one credit application at decision time (Model development document §3.1, generated fixture), and the evidence behind every figure below is 12,000 scored record(s), one credit application at decision time as at 2026-08-31.

WHAT THIS DOSSIER IS NOT: it is a validator's record of what was measured and what was not. It is not a conformity assessment, it does not issue the declaration of conformity that section 8 asks about, and nothing in it establishes that this system may be placed on the market.

No third-party dependency is declared for this engagement. That is a statement by the institution that the model embeds no vendor or open-source component it cannot validate inside — not an absence of the question.

NOT PERFORMED — AND WHY

The Annex IV technical documentation dossier: system description, development process, and the metrics used to measure accuracy

Required by EU AI Act (Regulation (EU) 2024/1689), Article 11 and Annex IV. Established by the dossier itself, recorded as evidence with its hash — the same document the declared accuracy metric in this pack's AUC row is cited to — a record the provider holds. No measurement over a scored population evidences it, and this dossier reports whether it is on file rather than standing in for it.

2 Development, data governance, and the metrics used to measure accuracy

Annex IV(2) asks how the system was developed, where its training, validation and testing data came from, and WHICH METRICS THE PROVIDER DECLARED to measure accuracy and robustness. This platform did not develop the system and holds none of its development record; what it holds is the declaration the provider made about the book it supplied, and the figures computed against it.

The declaration under test: records are identified by "application_id" and dated by "decision_date"; scores are banded at 0, 0.1, 0.25, 0.5, 1; the population is split by "segment"; the operating cutoff is 0.35; and the reference window for stability ends 2026-03-31. THE SAME BYTES READ AGAINST A DIFFERENT BANDING OR A DIFFERENT WINDOW ARE A DIFFERENT RUN, not a better-formatted one, which is why the declaration is stored beside the book and hashed with it.

NO ACCURACY CLAIM IS DECLARED for this system. Article 13(3)(b)(ii) and Article 15(3) require the instructions for use to state the declared levels of accuracy and the metrics they are measured by; with none declared, every figure in section 3 is reported and none of them is judged. That is a statement about the documentation, not about the model.

NO ATTRIBUTION ARTEFACT RESOLVED for this system. A machine-learning model has no declarable specification, so an attribution export is the only window onto what it relies on; without one, this dossier can report how the system performed and cannot report what it performed on.

NOT PERFORMED — AND WHY

Instructions for use stating the declared levels of accuracy and the relevant accuracy metrics

Required by EU AI Act, Article 13(3)(b)(ii) and Article 15(3). Established by the instructions for use, recorded as evidence — the source document of the accuracy claim, without which this pack's AUC row is reported unjudged — a record the provider holds. No measurement over a scored population evidences it, and this dossier reports whether it is on file rather than standing in for it.

Data governance for the training, validation and testing sets: provenance, preparation, and examination for bias

Required by EU AI Act, Article 10. Established by the data governance record, recorded as evidence. A scored population evidences what the model OUTPUT, never what it was trained on: the training set is not in this book and no figure over this book can stand in for one — a record the provider holds. No measurement over a scored population evidences it, and this dossier reports whether it is on file rather than standing in for it.

The human oversight measures built into the system, and the competence of those exercising them

Required by EU AI Act, Article 14. Established by the oversight procedure and the record of who exercises it, recorded as evidence. Whether an override happened is in the log; whether the oversight is meaningful is a reading of the procedure — a record the provider holds. No measurement over a scored population evidences it, and this dossier reports whether it is on file rather than standing in for it.

3 Monitoring, functioning and control, including accuracy for specific groups

Over the whole population the system discriminates at 0.763, and no declared level of accuracy to judge it against. Its mean calibration bias is -116 bp.

Band by band — [0, 0.1): 1,124 record(s), +435 bp; [0.1, 0.25): 4,690 record(s), +154 bp; [0.25, 0.5): 4,749 record(s), -427 bp; [0.5, 1]: 1,437 record(s), -399 bp.

THE MEAN AND THE BANDS DISAGREE, AND THE BANDS ARE THE FINDING. Overall the model is under-predicting by 116 bp, which read alone is unremarkable; its worst band is out by 435 bp in the other direction. The two partly cancel in the mean, which is the failure mode of a model tuned on a rank metric and is invisible to a single calibration figure.

THE DEGREE OF ACCURACY FOR SPECIFIC GROUPS, which Annex IV(3) asks for by name — near-prime: 3,994 record(s), AUC 0.764; prime: 6,041 record(s), AUC 0.793; thin-file: 1,965 record(s), AUC 0.648. The system discriminates best in "prime" and worst in "thin-file", a spread of 0.145.

THAT TABLE IS REPORTED AND NOT GRADED. Whether a gap of this size is lawful, or acceptable for the decision this system feeds, is a legal and policy judgement about the affected population — Articles 10 and 27 put it with the provider and the deployer, and no supervisor publishes a figure a validator could test it against. This dossier's contribution is that the gap is measured, named and impossible to read past.

Against the declared reference window the score distribution moved by 0.1334 on the population stability index. The 0.10 it is compared against is an INDUSTRY CONVENTION and not a figure the regulation publishes. It is exceeded here, which says the distribution moved and does not say why: population change, an upstream data change and genuine regime change are not separable from a scored population, and naming one would be a guess under a heading.

The lowest-to-highest selection rate across declared segments at the declared operating cutoff is 0.45. IT IS REPORTED AND PERMANENTLY UNJUDGED, and this dossier says so at more length than its US counterpart because of who reads it. A reader here may take a selection-rate gap for an Article 10(2)(f) bias finding; it is not one. The four-fifths rule that makes 0.80 feel like a threshold is 29 CFR 1607.4(D), a US employment-selection guideline with no standing under this regulation or in model supervision anywhere. The AI Act sets no ratio. Shipping one as a bar would be this platform asserting a legal standard that does not exist, and the ratio this system must hold is the provider's own declared policy.

Against the findings register: 2 finding(s) open, 1 at grade 1, 2 carrying a grade the engine PROPOSED and nobody has owned, 2 newly raised. Oldest open finding is 0 day(s) old.

2 live finding(s) have no owner assigned. An unowned finding has no one accountable for closing it and will age without anyone noticing.

Out-of-sample discrimination (AUC) on the scored population**MEASURED 0.763**

Out-of-sample discrimination (AUC) on the scored population: 0.763 measured on 12,000 scored records carrying 3,605 realised positive outcomes. Benchmark source: one credit application at decision time (Model development document §3.1, generated fixture), as at 2026-08-31.

Calibration bias (mean predicted probability – realised outcome rate)**MEASURED -116 bp**

Calibration bias (mean predicted probability – realised outcome rate): -116 bp measured on 12,000 scored records carrying 3,605 realised positive outcomes. Benchmark source: one credit application at decision time (Model development document §3.1, generated fixture), as at 2026-08-31.

Worst declared score band: absolute calibration bias in that band**MEASURED 435 bp**

Worst declared score band: absolute calibration bias in that band: 435 bp measured on 12,000 scored records carrying 3,605 realised positive outcomes. Benchmark source: one credit application at decision time (Model development document §3.1, generated fixture), as at 2026-08-31.

Score stability (PSI) between the declared reference window and everything after it**MEASURED 0.1334**

Score stability (PSI) between the declared reference window and everything after it: 0.1334 measured on 12,000 scored records carrying 3,605 realised positive outcomes. Benchmark source: one credit application at decision time (Model development document §3.1, generated fixture), as at 2026-08-31.

Segment performance spread (best minus worst declared segment, AUC)**MEASURED 0.145**

Segment performance spread (best minus worst declared segment, AUC): 0.145 measured on 12,000 scored records carrying 3,605 realised positive outcomes. Benchmark source: one credit application at decision time (Model development document §3.1, generated fixture), as at 2026-08-31.

Selection-rate ratio between declared segments at the declared operating cutoff**MEASURED 0.45**

Selection-rate ratio between declared segments at the declared operating cutoff: 0.45 measured on 12,000 scored records carrying 3,605 realised positive outcomes. Benchmark source: one credit application at decision time (Model development document §3.1, generated fixture), as at 2026-08-31.

MATERIAL ml-band-cancellation Calibration errors in opposite directions cancel in the overall bias

The overall calibration bias and the worst declared band's bias point in opposite directions, and the band's is more than twice the size. A reader of the headline figure alone would conclude the model is close to calibrated; band by band it is over-confident at one end of the score range and under at the other.

Evidence to re-run: Band table over 12,000 scored record(s), one credit application at decision time, against the score bands declared for this engagement.

NOTABLE ml-segment-spread One declared segment is served materially worse than the population as a whole

Discrimination in "thin-file" is below the population figure by more than five points of AUC. This is a finding about the MODEL and not a fairness conclusion: whether a gap of this size is lawful or acceptable is a judgement about the population and the decision it feeds, which this review does not make.

Evidence to re-run: Per-segment AUC over the declared segment column, 12,000 scored record(s), one credit application at decision time.

NOT PERFORMED — AND WHY

Independent replication of the model, and every check resting on a replica

PERMANENT, on this platform's ground. A gradient-boosted machine, a random forest and a neural network are not expressible as a declared specification — there are no coefficients to state — so no replica can be built and the replication fidelity gate has nothing to run. This is a property of how the model was built and NO UPLOAD CHANGES IT; the capability ruling says so with a permanent gate rather than a pending one. A replica fitted to the model's input/output behaviour instead would agree with it exactly where it was sampled and diverge everywhere else, and the fidelity gate would pass because the fixture is where it looks

Anything requiring the model to be re-run: robustness under perturbation, adversarial probing, counterfactual sweeps

These need the model executed on inputs this platform generates, which is the runtime capability — permanently gated, not pending. Every figure in this pack is computed from outputs the model already produced. Perturbation robustness is the most-requested of the three and the most misleading to approximate: perturbing the SCORED POPULATION rather than the model measures the arithmetic of this pack, not the stability of the model

A fairness verdict — whether the selection-rate ratio between segments is acceptable

The ratio is REPORTED at the declared cutoff and permanently unjudged. The four-fifths rule is 29 CFR 1607.4(D), a US employment-selection guideline with no standing in model supervision, and shipping it as a bar would be this product asserting a legal standard in a domain it does not come from. Fraud's decision, unchanged. What a model must hold is the institution's own policy, declared as a claim

Explainability computed by this platform — SHAP values, attributions, partial dependence

Explainability artefacts are DECLARED EVIDENCE: supplied by the owner, hashed, tenant-scoped and re-verified on read, exactly like every other artefact here. This product will not compute them from a replica it authored, and that is the load-bearing refusal — computing an explanation of a model we reconstructed and then validating the reconstruction against itself is the independence hole this platform closed for specifications, arriving one artefact later. It would also be impossible here: there is no replica, which is the premise of this whole family

Any reading of feature importance as a statement about the world

An attribution is a statement about a MODEL and never about the world. A memo that let one read as the second would be making a causal claim nobody evidenced — that a feature DRIVES an outcome rather than that a model USES it — and the two are routinely confused in exactly the documents this platform reads

Why the score distribution moved

That it moved is measurable and is this pack's PSI row. WHY — population change, an upstream data change, or genuine regime change — is not recoverable from a scored population, and the three call for different action. A pack that named a cause would be guessing under a heading, and the guess would be acted on

Multi-class, ranking and regression models

This pack reads a SCORED POPULATION — a score per record and a realised binary outcome — which is what an adverse-action or eligibility model produces. A regressor's error distribution, a ranker's ordering quality and a multi-class model's per-class confusion have nothing in common with it or with each other, and a pack covering all of them would compute their intersection, which is almost nothing. Named here rather than silently mis-served: a second pack, with its own scoping

4 Appropriateness of the performance metrics

Annex IV(4) asks the provider to justify that the metrics used are the appropriate ones for this system. THIS SECTION RECORDS THAT REASONING AND DOES NOT SUPPLY IT. A validator can say what was computed, over which population and against which declared claim; a validator cannot certify that discrimination was the right question to ask of this model, because that follows from what the system decides and for whom — facts the provider holds and this review does not.

What was computed: Out-of-sample discrimination (AUC) on the scored population; Calibration bias (mean predicted probability – realised outcome rate); Worst declared score band: absolute calibration bias in that band; Score stability (PSI) between the declared reference window and everything after it; Segment performance spread (best minus worst declared segment, AUC); Selection-rate ratio between declared segments at the declared operating cutoff.

WHAT THE PROVIDER DECLARED: nothing. With no declared level of accuracy on file, the question this section exists to record has no answer in the record, and the figures above stand as measurements with no stated purpose behind them.

Two of this pack's six rows can never be judged by anyone: the population stability index is compared against an industry convention, and the selection-rate ratio against nothing at all. Their presence in section 3 is not a claim that they were tested.

NOT PERFORMED — AND WHY

The provider's justification that these are the appropriate metrics

Annex IV(4) asks for it and no declared accuracy level is on file, so there is nothing to record. The absence is stated here rather than filled with this platform's own view of which metric suits this system.

5 Risk management system (Article 9)

Article 9 requires a risk management system running across the whole lifecycle, with the residual risks judged acceptable and communicated. NOTHING MEASURED OVER A SCORED POPULATION EVIDENCES THAT SUCH A PROCESS EXISTS. This section records which risk management record is on file and what it says; it does not infer one from the figures in section 3.

NOT PERFORMED — AND WHY

A risk management system running across the whole lifecycle, with the residual risks judged acceptable and communicated

Required by EU AI Act, Article 9; NIST AI RMF 1.0, GOVERN and MANAGE. Established by the risk management record, recorded as evidence. NIST AI RMF is a VOLUNTARY framework and explicitly not a conformity standard: it publishes no thresholds, and a template citing it as though it set a bar would misrepresent it — a record the provider holds. No measurement over a scored population evidences it, and this dossier reports whether it is on file rather than standing in for it.

6 Relevant changes made through the system's lifecycle

A validation run is a statement about ONE version of one system, identified in section 1. Whether the system has since moved — a retrain, a threshold change, a new feature, a new data source — is a fact only the provider holds, and this dossier neither knows nor can detect it.

NOT PERFORMED — AND WHY

The record of relevant changes through the system's lifecycle

Held by the provider. This review sees one book, scored by one version, and a change made after that book was cut would leave every figure here true of a system that no longer exists.

7 Harmonised standards applied, or the alternative solutions adopted

Which harmonised standards were applied, and where none was applied, what was adopted instead to meet the Chapter III requirements. This platform records the provider's list and ASSESSES CONFORMITY TO NONE OF THEM: conformity to a harmonised standard is an assessment against that standard's own text, which is not what any figure in section 3 measures.

NOT PERFORMED — AND WHY

Conformity to any harmonised standard

Not assessed, and not assessable from a scored population. The presumption of conformity a harmonised standard confers follows from an assessment against the standard, not from a validation of the model's outputs.

8 EU declaration of conformity

THIS PLATFORM CANNOT PRODUCE AN EU DECLARATION OF CONFORMITY AND WILL NOT STAND IN FOR ONE. The declaration is the provider's own legal instrument, drawn up and signed under their liability, asserting that the system meets Chapter III. No validator issues it, no evidence generates it, and no certificate this platform issues is one.

This section records whether the declaration has been drawn up and by whom. An absence here is stated and left as an absence.

NOT PERFORMED — AND WHY

The EU declaration of conformity

Not produced by this review and not producible by it. Anything issued here that looked like one would be a validator asserting a conformity only the provider can assert.

9 The post-market monitoring plan and what it has reported (Article 72)

Article 72 requires a post-market monitoring plan and reporting against it. The stability figure in section 3 is ONE PERIOD of the kind of evidence such a plan collects — a single comparison of the score distribution against one declared reference window. That a plan exists, at what interval it runs, who reads it and what it has changed is governance, evidenced by the plan itself and by its reports.

NOT PERFORMED — AND WHY

The post-market monitoring plan and what it has reported since deployment

Required by EU AI Act, Article 72. Established by the monitoring plan and its most recent report, recorded as evidence. This pack's PSI row is one period's evidence of the kind such a plan collects; that the plan exists, at what interval it runs and what it changed is governance — a record the provider holds. No measurement over a scored population evidences it, and this dossier reports whether it is on file rather than standing in for it.

10 Coverage of this dossier, and what it does not evidence

EU AI Act — Technical Documentation (EU · Annex IV) covers 3 model family(ies). 1 of them is UNDECLARED — nothing is measured, nothing is decided, and this dossier carries no evidence at all for simulation. That is stated here rather than left to be inferred from the sections it was able to fill. genai is evaluated by a harness that deliberately has no metric pack; those figures carry their own reporting basis and are not a pack's. 1 family(ies) have a pack behind them, and every figure in this document is still judged only against a published instrument or the institution's own declared claim.

simulation — UNDECLARED — NOTHING IS MEASURED HERE. No metric pack, no evaluation harness and no decided scope exists for the "simulation" family. This is UNDECLARED rather than "not applicable": nobody has ruled that this regime's requirements do not reach it, and a dossier that passed over it in silence would read as though somebody had. Any part of this document that would rest on measured evidence for "simulation" rests on nothing.

genai — EVALUATED WITHOUT A PACK. There is NO METRIC PACK for GenAI and that is a declared position rather than a gap: SR 26-2 puts GenAI outside its own scope, so a pack here would assert a supervisory requirement no supervisor has made. What this platform runs instead is an evaluation harness — trace book, judge calibration, groundedness, injection resistance and tool-scope conformance — whose figures are reported with their own basis and which produces no memo and no certificate. A dossier carries those figures and must not present them as a pack's.

ml — COMPUTED. The "ml" pack computes 6 metric(s) for this family from what the model DID. It computes nothing from a replica, and none ever will exist: Machine-learning models in this family are gradient-boosted trees, random forests and neural networks. None of them is expressible as a declared specification — there are no coefficients to state — so no replica can be built and the replication fidelity gate has nothing to run. This is a property of how the model was built and no upload changes it. What this platform does perform on such a model is its outcomes, its calibration, its stability and its operating points, plus the evidence artefacts its owner can supply.

NOT PERFORMED — AND WHY

genai — no computed evidence in this dossier

There is NO METRIC PACK for GenAI and that is a declared position rather than a gap: SR 26-2 puts GenAI outside its own scope, so a pack here would assert a supervisory requirement no supervisor has made. What this platform runs instead is an evaluation harness — trace book, judge calibration, groundedness, injection resistance and tool-scope conformance — whose figures are reported with their own basis and which produces no memo and no certificate. A dossier carries those figures and must not present them as a pack's.

simulation — no computed evidence in this dossier

No metric pack, no evaluation harness and no decided scope exists for the "simulation" family. This is UNDECLARED rather than "not applicable": nobody has ruled that this regime's requirements do not reach it, and a dossier that passed over it in silence would read as though somebody had. Any part of this document that would rest on measured evidence for "simulation" rests on nothing.



AlphaIQ CertifyIQ — independent validation of models and the decisions they drive. This sample was produced by the product itself; nothing in it was written by hand.

Memo content hash fde7c63518038aedd868069c79a739f7d88c091ea33e45a20cefd63e8366eccc

This document is a recommendation, not a certification. A certificate is issued only on attested evidence, which a synthetic sample does not have.