

Q-RAD: a gated residual architecture for quantum-assisted fraud detection, and why its quantum branch does not pay

Jean Marie Vianney Sindayigaya

Abstract—Hybrid quantum–classical detectors are usually evaluated as monolithic alternatives to classical models, which conflates two separable questions: whether the hard cases can be identified, and whether a quantum model can resolve them. We introduce Q-RAD, a gated residual architecture that separates these questions by construction. A quantum-blind, quantum-cost-free gating function routes only ambiguous transactions to a financial-transaction-specific quantum-kernel branch, whose scores are fused residually with a classical trunk under an explicit finite-shot budget. We predeclare the architecture, an eight-comparison confirmatory family with Holm–Bonferroni correction, a cost metric in quantum circuit executions, and a refutation posture, then evaluate on the Kaggle Credit-Card Fraud dataset over ten repeated case-control splits under three blocking validity gates, including exact-equality replay of previously published results. The routing half of the architecture succeeds: the gate attains 0.889 AUROC against realized classical errors and captures 77% of them within a 22% routing budget. The quantum half fails: handed exactly those hard cases, the quantum residual produced zero net corrections in nine of ten seeds and one net regression in the tenth. Q-RAD attains mean case-control test F1 0.916 against 0.918 for the classical trunk and 0.922 for an otherwise identical classical second opinion; the hypotheses of improvement over the classical trunk, over random routing, and over a classical second opinion are all refuted after correction. We report this refutation as predeclared, and localise it: ambiguity in this task is identifiable from classical signals alone, but the quantum kernel cannot exploit that signal at any tested shot budget. Balanced-cohort metrics are benchmark values, not deployment-prevalence estimates.

Keywords—Quantum Machine Learning; Quantum Kernel Methods; Selective Prediction; Learning to Defer; Hybrid Quantum-Classical Models; Financial Fraud Detection; Shot Noise; Negative Results

I. INTRODUCTION

QUANTUM kernel methods are routinely proposed for financial fraud detection, and are routinely evaluated in the same way: a quantum model is trained on the whole dataset, a classical model is trained on the whole dataset, and their aggregate scores are compared [1], [2]. This framing conflates two questions that are separable in principle and very different in practice. First, *can the hard cases be identified* — is there a usable signal, available before any quantum computation is performed, that marks the transactions a classical

detector will get wrong? Second, *can a quantum model resolve them* — given exactly those transactions, does a quantum feature map supply discriminative structure the classical model lacked? A monolithic comparison answers neither question; it returns a single number in which both effects are entangled with each other and with the cost of running a quantum device on every input.

This paper separates the two questions architecturally. We introduce *Q-RAD*, the Quantum Residual Anomaly Detector: a classical trunk, a learned gating function that scores each transaction’s ambiguity, and a quantum-kernel branch that is consulted only for the routed minority, whose score is fused residually with the trunk’s. The gate is *quantum-blind* — it never observes any quantum output on any data split — and *quantum-cost-free* — routing decisions consume no circuit executions. This makes the cost model non-circular and lets each half of the architecture be refuted independently.

The design descends from the reject-option and selective-prediction literature [3]–[5], where a model may abstain on inputs it is likely to misclassify, and from learning to defer [6], where abstention hands the case to a second decision maker. Q-RAD’s second decision maker is a quantum kernel whose per-call cost is explicit rather than assumed free, which is the operationally relevant regime for near-term devices [7]–[9].

The quantum branch is FT-FM-lite, the tunable-entanglement refinement of the financial-transaction feature map FT-FM introduced in the first two papers of this series [10], [11]. That prior work established the honest baseline this paper must respect: on the Kaggle Credit-Card Fraud dataset under a leakage-controlled case-control protocol, the quantum kernel attained mean test F1 0.891 against 0.918 for a matched classical RBF-SVM, with the classical model significantly ahead after multiplicity correction. Q-RAD’s bet is therefore narrow and falsifiable: the two models’ errors need not coincide, so routing plus residual fusion might beat both pure strategies at a fraction of the quantum cost — or might not, and the protocol must be able to say so.

We report that it does not. The routing half of the architecture works well; the quantum half contributes nothing. Our contributions are as follows. 1) We define Q-RAD and its quantum-blind, quantum-cost-free gate, together with a residual fusion rule whose zero-strength endpoint provably

Author is with Warsaw University of Technology, Poland (e-mail: jean_marie_vianney.sindayigaya.dokt@pw.edu.pl).



reproduces the classical trunk (Section III). 2) We predeclare a protocol with three blocking validity gates — including exact-equality replay of the previously published quantum result — an eight-comparison confirmatory family under one Holm–Bonferroni correction, explicit small-sample power arithmetic, and a cost metric denominated in quantum circuit executions (Section IV). 3) We report the canonical result: the gate attains 0.889 AUROC against realized trunk errors and captures 77% of them within a 22% routing budget, while the quantum residual produces zero net corrections in nine of ten seeds (Section V). 4) We publish the refutation as predeclared and localise its cause to the quantum branch rather than to the architecture, which we argue is the more useful negative result for the field (Section VI).

The remainder of the paper is organised as follows. Section II reviews selective prediction and quantum kernels under finite shots. Section III defines Q-RAD. Section IV gives the predeclared protocol. Section V reports results, Section VI discusses them and bounds the claims, and Section VII concludes.

II. BACKGROUND AND RELATED WORK

A. Selective prediction and learning to defer

A classifier with a reject option may decline to decide on inputs where its expected loss is high, trading coverage for accuracy along a curve first characterised by Chow [3]. Modern selective classification learns the abstention rule jointly with, or on top of, the predictor and evaluates it by risk-coverage analysis [4]; simple confidence functionals such as the maximum softmax probability remain strong baselines [5], and distribution-free coverage guarantees are available through conformal methods [12]. Learning to defer extends abstention to a genuine hand-off: the rejected case is routed to a second decision maker, and the gate is trained against the composite system’s loss rather than the base model’s alone [6].

Q-RAD is a learning-to-defer system whose deferral target is a quantum kernel. Two design consequences follow. First, because the deferral target carries a hardware cost that the base model does not, the gate must be trainable and evaluable without consulting it — otherwise the cost accounting becomes circular. Second, because the quantum branch is the object under study, the gate must never observe its outputs, or an apparent architectural gain could be gate overfitting to quantum idiosyncrasies rather than genuine complementarity.

B. Quantum kernels, concentration, and finite shots

A quantum feature map $|\psi(x)\rangle = U(x)|0\rangle^{\otimes n}$ induces the fidelity kernel $K(x, x') = |\langle\psi(x) | \psi(x')\rangle|^2$, which a support-vector classifier [13] can consume as a precomputed Gram matrix [1], [2]; indeed supervised quantum models are kernel methods in a precise sense [14]. The dominant failure mode is exponential concentration: for sufficiently expressive maps, kernel values concentrate around a fixed value independently of input similarity, leaving the Gram matrix uninformative [15], an effect connected to the expressibility–trainability tension of variational models [9], [16]. Global rescaling of encoding angles — kernel bandwidth — is one

control that moves a model along this curve [17]; FT-FM-lite’s entanglement-strength parameter β is an analogous control acting on the entangling interactions only [11].

Estimating K on hardware is a sampling problem. The compute-uncompute circuit yields the all-zeros outcome with probability $K(x, x')$, so a B -shot estimate is a binomial frequency: unbiased per entry with variance at most $1/(4B)$, but with $O(1/B)$ bias in every nonlinear statistic computed downstream [7]. The number of entries required scales quadratically in the training-set size, which is the practical bottleneck for quantum support-vector machines [8]. Prior quantum anomaly detection work has largely assumed exact kernels [18]; we instead treat the shot budget as a first-class experimental variable, because a gated architecture’s entire value proposition is expressed in circuit executions saved.

C. Positioning

Against this background, the contribution of this paper is neither a new feature map nor a new kernel estimator, but an evaluation architecture that makes the two halves of the hybrid claim separately falsifiable, applied under a protocol strict enough that a negative result is informative rather than merely inconclusive. We are aware of no prior work that reports, for a financial fraud task, a gate quality measurement and a residual-correction count from the same run — the pair of measurements that localises where a hybrid detector fails.

III. THE Q-RAD ARCHITECTURE

A. Branches and residual fusion

Q-RAD comprises a classical trunk C , a quantum branch Q , and a gate R . The trunk is an RBF-kernel support-vector classifier [13] with balanced class weights, exposing a decision margin $m_c(x)$ and a Platt probability $p_c(x)$. The quantum branch is FT-FM-lite at frozen entanglement strength $\beta = 1$ [11] — the value selected in all ten seeds by the previous paper’s validation protocol — realised as a support-vector classifier over a shot-corrupted fidelity Gram matrix, exposing a margin $m_q(x)$.

Both margins are standardised by scale only,

$$z_b(x) = \frac{m_b(x)}{\sigma_b}, \quad b \in \{c, q\}, \quad (1)$$

where σ_b is the sample standard deviation of branch b ’s out-of-fold margins on the fitting cohort. There is no location shift, so $\text{sign } z_c = \text{sign } m_c$; the consequence is recorded as Proposition 1 below. The fused score and decision are

$$\begin{aligned} f(x) &= z_c(x) + \alpha R(x) (z_q(x) - z_c(x)), \\ \hat{y}(x) &= \mathbb{I}[f(x) \geq 0], \end{aligned} \quad (2)$$

with residual strength α selected on validation data from $\{0, 0.25, 0.5, 0.75, 1\}$ by maximum F1, ties broken by PR-AUC and then by smaller α .

Proposition 1 (collapse). If $\alpha = 0$, or if $R(x) = 0$ for all x , then $\hat{y}$ coincides with the trunk’s own decision rule. *Proof.* Both conditions reduce Eq. (2) to $f = z_c$, and by Eq. (1) with $\sigma_c > 0$ the sign of z_c equals the sign of m_c , which is the trunk’s decision function. $\square$

Proposition 1 matters for interpretation, not elegance: it guarantees that the $\alpha = 0$ grid cell lets the data select “no quantum”, so the architecture cannot be forced to consume quantum resources by construction, and any observed difference from the classical control is attributable to a deliberate, validated choice of $\alpha > 0$. It is verified against the deployed trunk by unit test.

B. The gate

The gate scores ambiguity with a low-capacity L_2 -regularised logistic model

$$a(x) = \sigma(\theta^\top [z_c(x), |z_c(x)|, p_c(x), \tilde{x}]), \quad (3)$$

where $\tilde{x}$ is the scaled feature vector, trained against *out-of-fold classical errors* on the fitting cohort. Routing is $R(x) = \mathbb{I}[a(x) \geq \tau]$, with τ the $(1 - \rho^*)$ -quantile of out-of-fold gate scores for a declared coverage target ρ^* . The gate is itself cross-fit, so τ is calibrated on genuinely held-out scores rather than resubstitution outputs. Two properties are enforced by construction:

- *Quantum blindness.* No input to Eq. (3), and no target used to fit θ , derives from any quantum computation on any split. The gate therefore cannot overfit to quantum idiosyncrasies, and its measured quality is a statement about the classical trunk’s error structure alone.
- *Cost freedom.* Evaluating $a(x)$ and thresholding it consumes zero circuit executions, so the cost saving claimed in Section V is not partially spent on deciding what to route.

A degenerate-gate guard falls back to the Chow margin heuristic $1 - |2p_c - 1|$ [3], [5] when the out-of-fold error target admits no stratified cross-fit or yields fewer than three distinct scores; the fallback is flagged per seed and per stage. It was not triggered in any seed of the canonical run.

C. Finite-shot kernel estimation

Each unique off-diagonal Gram entry is replaced by a binomial frequency

$$\hat{K}_{ij} \sim \text{Binomial}(B, K_{ij})/B, \quad (4)$$

emulating B executions of the compute-uncompute circuit; the diagonal is fixed at unity and square matrices are projected onto the positive semi-definite cone by eigenvalue clipping. Eq. (4) is unbiased per entry with variance $K_{ij}(1 - K_{ij})/B \leq 1/(4B)$. We claim nothing beyond the per-entry level: every nonlinear downstream statistic inherits $O(1/B)$ bias, and the unit diagonal together with the projection is a structured, disclosed distortion. At seven qubits, bulk fidelities concentrate near 2^{-7} [15], so small budgets zero most bulk entries and roughly 1.3×10^4 shots are needed for 10% relative error on a typical entry — an arithmetic reminder that exact-statevector results in the quantum kernel literature are optimistic ceilings.

D. Cost model

The headline cost metric is quantum circuit executions per classified transaction (QCE). For a fitting cohort of n rows, realized coverage ρ , and budget B , Q-RAD’s inference cost is ρnB against the always-quantum arm’s nB . Tuning is itemized separately and never hidden: the ledger records the Stage-A tuning Gram $\binom{240}{2} = 28\,680$ entries, the deployment Gram $\binom{320}{2} = 51\,040$ entries, the 80×240 validation selection rows, and the routed test rows. Because the classical trunk consumes no circuit executions at all, cost comparisons are staged only against the always-quantum arm at matched budget; no cost claim is made against the free classical baseline.

IV. PREDECLARED EXPERIMENTAL METHODOLOGY

Every element of this section was fixed in a written specification before any test cohort was evaluated, including the operating point, the comparison family, the claim rule, the cost thresholds, and the posture on failure.

A. Data and cohorts

We reuse the frozen protocol of the previous paper in this series [11] on the Kaggle Credit-Card Fraud dataset [19] so that results are comparable by construction. For each of ten seeds, raw rows are split 60/20/20 stratified before any transform is fitted; Time and Amount are retained and five of the PCA components V_1 – V_{28} are ranked by a training-only Fisher score, giving seven features within the qubit budget. Deterministic 1:1 case-control cohorts of 240/80/80 rows are drawn from the splits. Because every cohort is balanced by design, all reported F1 and PR-AUC values are case-control benchmark metrics and must not be read as deployment-prevalence estimates; the raw positive rate is $492/284,807 \approx 0.17\%$.

B. Stage order

Leakage control is enforced by stage ordering. Stage A fits the trunk, the quantum branch, and the gate on the 240-row training cohort. Stage B selects α on the 80-row validation cohort and freezes it together with ρ^* and β . Stage C refits every component on the 320-row train-plus-validation cohort and recalibrates τ from that cohort’s out-of-fold gate scores, so that the deployed gate audits the deployed trunk. Stage D evaluates the test cohort exactly once, purchasing quantum kernel rows only for routed points. All stochastic elements draw from a derived-seed namespace audited to be disjoint from the previous paper’s. Q-RAD and the always-quantum control share the identical noisy Gram and the identical test row draw at each budget, so their comparison uses common random numbers.

C. Operating point, controls, and ablations

The confirmatory operating point is $\rho^* = 0.25$, $B^* = 1024$, $\beta = 1$. Budgets $B \in \{256, 1024, 4096, \infty\}$ are evaluated, where ∞ denotes exact statevector simulation. Controls are: always-classical (the trunk’s own predictions); always-quantum at each budget; a random gate that routes a seeded uniform subset of the same size with the frozen α and

fusion rule; a *classical-residual second opinion* that replaces only the quantum branch with a seeded random forest [20] under an identical gate, coverage, and fusion rule; and the three additional classical baselines of the previous paper. All evaluations use noiseless statevector simulation with injected shot noise [21]; no hardware claim is made.

D. Blocking validity gates

Three gates must pass before any number is reported.

- 1) *Reproduction*. The $B = \infty$ always-quantum arm must reproduce the previously published per-seed F1 values under exact equality. This verifies that the pipeline reproduces the frozen contract bit-for-bit before anything new is added.
- 2) *Estimator equivalence*. The binomial injector of Eq. (4) is compared against genuine compute-uncompute fidelity sampling on a held-aside subset; every sampled pair must agree with the exact value within five standard deviations.
- 3) *Injector scaling*. The fitted slope of $\log\|\hat{K}_B - K\|_F$ on $\log B$ must lie in -0.5 ± 0.1 , as required by $B^{-1/2}$ Monte-Carlo scaling. Failure withholds the run entirely rather than a single budget, because a joint fit cannot attribute the failure to one budget.

E. Hypotheses, statistics, and power

The confirmatory family contains exactly eight two-sided Wilcoxon signed-rank tests [22] on paired per-seed F1, under a single Holm–Bonferroni correction at familywise $\alpha = 0.05$ [23]. Directionality is placed in the claim rule rather than the test: a *win* requires Holm significance *and* a median paired difference in the predicted direction, so significant losses are reported with the same prominence as wins.

- H1 Q-RAD > always-classical trunk at the operating point.
- H2 Q-RAD > always-quantum at B^* , under common random numbers.
- H3 Q-RAD > random gate at matched coverage.
- H7 Q-RAD > classical-residual second opinion.
- H5a–c Q-RAD versus random forest, logistic regression, and multilayer perceptron (two-sided, no predicted direction).
- H6 Always-quantum at $B = 256 <$ its exact-statevector value.

H4 is a non-inferential pass/fail cost check: mean realized coverage must lie in $[0.15, 0.35]$ and mean QCE must not exceed 0.35 times the always-quantum arm’s.

Small-sample power is predeclared rather than discovered. With $n = 10$ pairs the exact two-sided floor is $2^{-9} \approx 0.00195$, against a rank-one Holm bar of $0.05/8 = 0.00625$; attaining it requires at least nine effective pairs after zero differences are discarded. Because Proposition 1 makes exact ties likely whenever $\alpha = 0$ is selected, every comparison reports its effective pair count, its dropped-zero count, and its exact attainable p -floor, so a reader can verify whether each test was decidable at all. Routing headroom is likewise fixed in

TABLE I
HELD-OUT CASE-CONTROL TEST F1 OVER TEN REPEATED STRATIFIED SPLITS (MEAN $\pm$ STD). VALUES ARE BENCHMARK METRICS, NOT DEPLOYMENT-PREVALENCE ESTIMATES.

Arm	Budget B	Test F1
Classical second opinion (H7 control)	—	0.922 ± 0.030
Always-classical (trunk)	—	0.918 ± 0.032
Random gate	1024	0.918 ± 0.032
Q-RAD	1024	0.916 ± 0.032
Q-RAD	256	0.914 ± 0.037
Q-RAD	4096	0.918 ± 0.032
Q-RAD	∞	0.916 ± 0.031
Always-quantum	1024	0.881 ± 0.032
Always-quantum	256	0.879 ± 0.042
Always-quantum	∞	0.891 ± 0.031

advance: at $\rho^* = 0.25$ roughly twenty routed points face six or seven trunk errors per cohort, bounding the attainable effect to a few flipped decisions.

F. Refutation posture

We predeclare that failures of H1, H2, or H7 are published as refutations and not reframed; that non-significance is never reported as equivalence; and that the outcome “routing works, the quantum branch does not close the gap” is an acceptable and publishable result. Every reported number derives from a single archived artifact carrying its repository commit, a clean working-tree flag, pinned package versions, and the input file’s SHA-256 digest.

V. RESULTS

All values derive from a single archived run of ten seeds under the canonical Python 3.11 environment with a clean working tree. Balanced cohorts make every F1 a case-control benchmark metric, not a deployment-prevalence estimate.

A. Validity gates

All three blocking gates passed. The $B = \infty$ always-quantum arm reproduced the previously published per-seed F1 for all ten seeds under exact equality, confirming that the Q-RAD pipeline reproduces the frozen contract of the previous paper bit-for-bit. The estimator-equivalence check agreed with genuine compute-uncompute sampling to a maximum absolute deviation of 0.037 against a five-sigma tolerance of 0.078. The injector scaling slope was -0.499 ± 0.002 across seeds, against the required -0.5 ± 0.1 , with no failing seed. The degenerate-guard was not triggered in any seed.

B. Held-out performance

Table I and Fig. 1 report held-out F1 for every arm. Q-RAD attains mean case-control test F1 0.916, below the classical trunk’s 0.918 and below the classical-residual second opinion’s 0.922, and above the always-quantum arm’s 0.881 at the same budget. Validation selected $\alpha = 0$ in four of ten seeds and $\alpha \leq 0.25$ in nine.

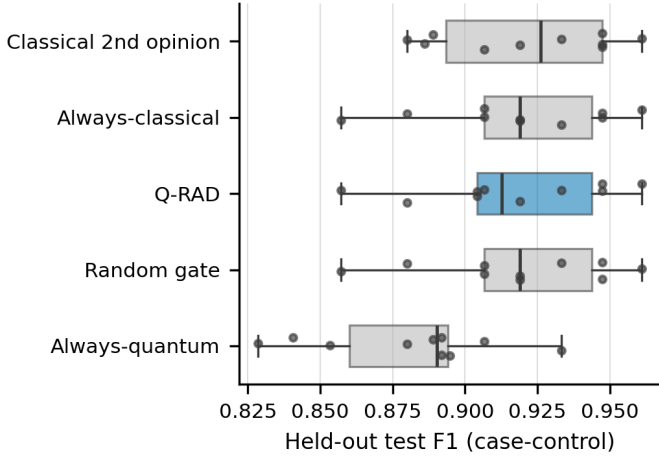


Fig. 1. Held-out case-control test F1 by arm at the predeclared operating point ($\rho^* = 0.25$, $B^* = 1024$), ten seeds each. Q-RAD sits between the classical arms and the always-quantum arm.

TABLE II

CONFIRMATORY FAMILY: TWO-SIDED WILCOXON SIGNED-RANK TESTS ON PAIRED PER-SEED F1, ONE HOLM-BONFERRONI CORRECTION, FAMILYWISE $\alpha = 0.05$. n_{EFF} IS THE PAIR COUNT AFTER ZERO DIFFERENCES ARE DISCARDED AND p_{min} THE EXACT ATTAINABLE TWO-SIDED FLOOR AT THAT COUNT.

	Control	Median $\Delta F1$	Raw p	Holm	n_{eff}
H1	Always-classical	0.000	0.500	No	2
H2	Always-quantum	+0.046	0.0059	Yes	10
H3	Random gate	0.000	0.500	No	2
H7	Classical 2nd opinion	0.000	0.500	No	3
H5a	Random forest	-0.002	0.641	No	8
H5b	Logistic regression	0.000	0.977	No	8
H5c	Multilayer perceptron	+0.067	0.0039	Yes	10
H6	Exact vs $B = 256$	-0.012	0.375	No	10

C. Confirmatory family

Table II reports the eight predeclared comparisons under one Holm-Bonferroni correction. H1, H3 and H7 are refuted: in each case the median paired difference is exactly zero and only two or three pairs survive the discarding of zero differences, giving exact attainable p -floors of 0.50 and 0.25 against a rank-one bar of 0.00625. These comparisons were therefore *structurally incapable* of significance, a consequence of Proposition 1: whenever validation selects $\alpha = 0$, Q-RAD’s predictions are identical to the trunk’s and the pair contributes no information. We report this as a limitation of the design at $n = 10$, not as evidence of equivalence.

The single Holm-significant win is H2, against the always-quantum arm, with median $\Delta F1 = +0.046$. We regard this as close to tautological rather than as evidence for the architecture: validation selection collapses toward the stronger classical arm, so outperforming the quantum-only arm is what the collapse mechanism of Proposition 1 guarantees. H6 is not significant: the $B = 256$ arm is lower than its exact-statevector reference in median as predicted, but the effect is not detectable at this sample size.

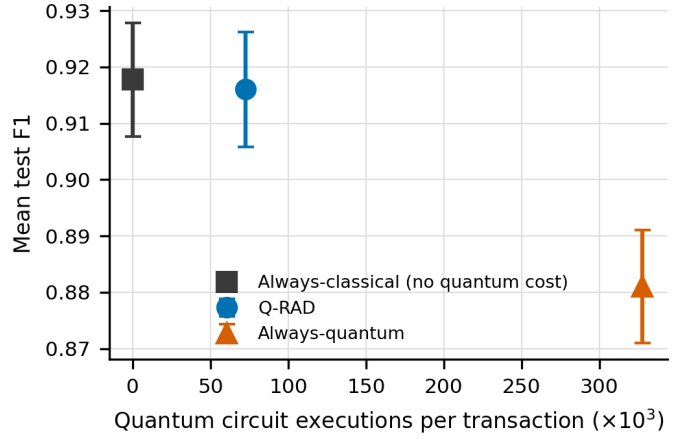


Fig. 2. Mean test F1 against quantum circuit executions per classified transaction at $B^* = 1024$; bars are standard errors over ten seeds. The classical trunk attains the highest accuracy at zero quantum cost.

D. Cost

H4 passed. Mean realized coverage was 0.221, inside the predeclared band $[0.15, 0.35]$, and mean QCE was 0.221 times the always-quantum arm’s, inside the 0.35 limit. Fig. 2 places the arms on the accuracy–cost plane and makes the operational conclusion immediate: the always-classical arm occupies the top-left corner, attaining the highest mean F1 at exactly zero quantum cost, so no cost argument can rescue either quantum-consuming arm on this task.

E. Mechanism: where the architecture succeeds and where it fails

The mechanism diagnostics in Fig. 3 separate the two halves of the hybrid claim, which is the purpose for which the architecture was built.

The gate succeeds. Across seeds it attains mean AUROC 0.889 against the trunk’s realized test errors and captures 77.1% of those errors inside a 22% routing budget, with mean coverage drift $|\rho - \rho^*| = 0.061$. Ambiguity in this task is therefore identifiable in advance from classical signals alone, at no quantum cost.

The quantum residual fails. Given precisely those correctly identified hard cases, the net number of decisions it flipped from wrong to right, minus those flipped from right to wrong, was zero in nine of ten seeds and -1 in the tenth. The routed transactions are genuinely the ones the trunk gets wrong, and the quantum kernel does not resolve them.

F. Shot-budget behaviour

Fig. 4 shows mean F1 against budget for both quantum-consuming arms. Neither is monotone in B : mean F1 at $B = 4096$ slightly exceeds the exact-statevector reference for both arms. We report this as a predeclared anomaly diagnostic rather than a finding; the departures are of the order of a single flipped decision per cohort and well inside the seed-to-seed spread. The always-classical level is shown for reference and is not exceeded by either arm at any budget.

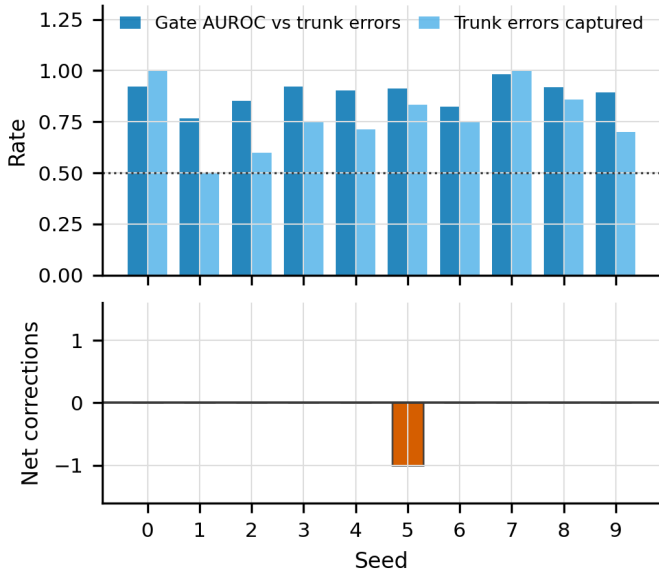


Fig. 3. Top: gate quality per seed — AUROC against realized trunk errors, and the fraction of those errors captured inside the 22% routing budget. Bottom: net corrections contributed by the quantum residual, zero in nine seeds and -1 in one (highlighted).

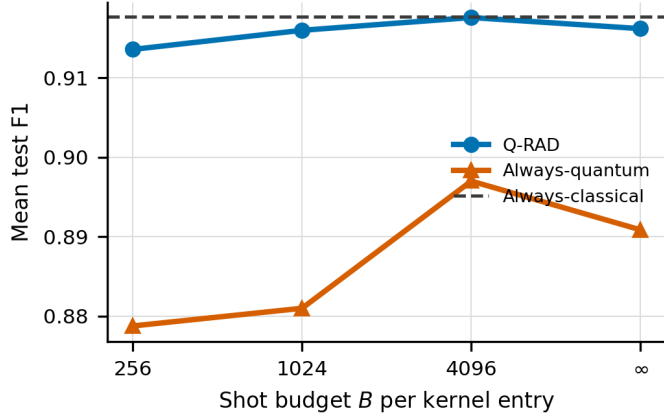


Fig. 4. Mean test F1 against shot budget B per kernel entry. Neither quantum-consuming arm reaches the classical trunk at any budget, including exact statevector simulation.

VI. DISCUSSION

A. What was refuted, and what was not

The predeclared bet had two halves. The routing half — that the transactions a classical fraud detector gets wrong can be identified in advance, cheaply, from classical signals alone — is supported: AUROC 0.889 against realized errors, 77% of them captured within a 22% budget, no fallback triggered in any seed. The residual half — that an FT-FM-lite quantum kernel can resolve those cases — is refuted: zero net corrections in nine of ten seeds, one net regression in the tenth, and no Holm-significant improvement over the classical trunk, over random routing, or over a classical second opinion.

The value of separating the two halves is precisely that the refutation is *localized*. A monolithic comparison of Q-RAD against the trunk would have produced 0.916 versus 0.918 and licensed only the vague conclusion that the hybrid did

not help. The gate-quality and net-correction measurements together identify the failing component: not the selective-prediction logic, not the gate’s feature set, not the coverage calibration, but the quantum branch’s ability to discriminate on the subpopulation it was given.

We resist one tempting over-reading. Failing to reject the null is not evidence of equivalence, and here it is not even evidence of a small effect: H1, H3 and H7 retained only two or three effective pairs, with attainable p -floors above the Holm bar, so those tests could not have succeeded whatever the underlying truth. The substantive evidence against the quantum residual is not their p -values but the direct mechanism measurement, which does not depend on the significance machinery at all.

B. Why the quantum branch may have failed here

Three explanations are consistent with the evidence and are not separated by this experiment. First, concentration: at seven qubits with $\beta = 1$ the FT-FM-lite kernel is known to be strongly concentrated [11], and concentrated kernels carry little usable similarity structure [15]. Second, subpopulation difficulty: the routed cases are by construction those on which a strong classical model already fails, and they may be intrinsically ambiguous — label noise or genuine overlap — rather than merely misaligned with the classical kernel; the oracle ceiling on this subpopulation was not estimated in a way that distinguishes these. Third, representational overlap: the quantum branch sees the same seven PCA-derived features as the trunk, so its errors may simply correlate with the trunk’s, which is exactly what the near-zero net-correction count would look like. Distinguishing these is the natural next experiment and would be more informative than another end-to-end hybrid comparison.

C. Implications for hybrid evaluation practice

Two elements of this protocol are transferable independently of the result. The first is the classical-residual second opinion: an otherwise identical architecture in which only the quantum branch is replaced by a classical model. Without it, any gain observed in a gated hybrid may be attributable to routing and ensembling rather than to quantum computation, and here the control in fact *outperformed* the quantum version. We suggest such a control should be standard in gated hybrid studies.

The second is cost-denominated reporting. Because the gate is quantum-blind and quantum-cost-free, quantum circuit executions per transaction is a non-circular metric, and the accuracy–cost plane of Fig. 2 makes the conclusion visible at a glance: an arm at zero quantum cost dominates. We also note that reporting effective pair counts alongside p -values changed the interpretation of this paper materially, and costs nothing to adopt.

D. Limitations and threats to validity

Seven limitations bound these claims. First, the ten seeds are repeated stratified splits of a single dataset, so intervals quantify protocol stability rather than inference across independent populations. Second, all cohorts are balanced by

design, so every metric is a case-control benchmark value and not a deployment-prevalence estimate. Third, the effective-pair counts above mean three confirmatory comparisons were undecidable at this sample size; larger seed counts or an operating point that forces $\alpha > 0$ would be needed to resolve them. Fourth, shot noise is injected into exact statevector kernels rather than sampled on hardware, so device noise, connectivity, and calibration drift are entirely out of scope and no hardware claim is made. Fifth, the quantum branch is frozen at $\beta = 1$ inherited from the previous paper; a jointly tuned β within the gated architecture was not explored. Sixth, the gate family, coverage target, and fusion rule were fixed in advance and only the reported ablations vary them, so a better gated architecture may exist. Seventh, results on one financial dataset do not establish generality; the network-intrusion datasets named in the first paper of this series remain deferred pending frozen column contracts.

E. Reproducibility

The canonical artifact records the repository commit, a clean working-tree flag, pinned package versions, the input file's SHA-256 digest, per-seed source-row indices, the full derived-seed namespace, itemized execution ledgers, and all three gate reports. Every number in Section V is traceable to that artifact, and no number in this paper was produced by any other means.

VII. CONCLUSION

We introduced Q-RAD, a gated residual architecture in which a quantum-blind, quantum-cost-free gating function routes only ambiguous transactions to a financial-transaction-specific quantum-kernel branch under an explicit finite-shot budget, and we evaluated it under a protocol predeclared down to the operating point, the eight-comparison confirmatory family, the cost thresholds, the small-sample power arithmetic, and the posture on failure.

The architecture separates two questions that monolithic hybrid comparisons conflate, and the answers diverge. Ambiguity is identifiable: the gate reached 0.889 AUROC against realized classical errors and captured 77% of them within a 22% routing budget, at no quantum cost. The quantum residual is not useful: handed exactly those cases, it produced zero net corrections in nine of ten seeds and one net regression in the tenth. Q-RAD attained mean case-control test F1 0.916 against 0.918 for the classical trunk and 0.922 for an otherwise identical classical second opinion; the hypotheses of improvement over the trunk, over random routing, and over a classical second opinion are refuted after multiplicity correction. The cost check passed — Q-RAD used 22% of the always-quantum arm's circuit executions — but no cost claim can be staged against a classical baseline that consumes none. We claim no quantum advantage on this task at any tested shot budget, including exact statevector simulation.

We report this refutation as predeclared, and we regard its localisation as the contribution. The failure lies in the quantum branch, not in the selective-prediction architecture around it, and the two protocol elements that establish this

— a classical-residual second-opinion control, and mechanism measurements reported alongside aggregate scores — transfer to any gated hybrid study independently of the outcome.

The next paper in this series develops quantum-native explainability for the FT-FM family using entanglement entropy and reduced density matrices. The present result sharpens its motivation: before a quantum module can be made auditable for regulated deployment, there must be evidence that it contributes decisions worth auditing, and establishing that contribution — on data where the classical baseline is strong, under honest cost accounting — remains open.

REFERENCES

- [1] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, "Supervised learning with quantum-enhanced feature spaces," *Nature*, vol. 567, pp. 209–212, 2019, <https://doi.org/10.1038/s41586-019-0980-2>. [Online]. Available: <https://doi.org/10.1038/s41586-019-0980-2>
- [2] M. Schuld and N. Killoran, "Quantum machine learning in feature hilbert spaces," *Physical Review Letters*, vol. 122, no. 4, p. 040504, 2019, <https://doi.org/10.1103/PhysRevLett.122.040504>. [Online]. Available: <https://doi.org/10.1103/PhysRevLett.122.040504>
- [3] C. K. Chow, "On optimum recognition error and reject tradeoff," *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 41–46, 1970, <https://doi.org/10.1109/TIT.1970.1054406>. [Online]. Available: <https://doi.org/10.1109/TIT.1970.1054406>
- [4] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," 2017, advances in Neural Information Processing Systems 30. <https://doi.org/10.48550/arXiv.1705.08500>. [Online]. Available: <https://doi.org/10.48550/arXiv.1705.08500>
- [5] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," 2017, international Conference on Learning Representations. <https://doi.org/10.48550/arXiv.1610.02136>. [Online]. Available: <https://doi.org/10.48550/arXiv.1610.02136>
- [6] D. Madras, T. Pitassi, and R. Zemel, "Predict responsibly: Improving fairness and accuracy by learning to defer," 2018, advances in Neural Information Processing Systems 31. <https://doi.org/10.48550/arXiv.1711.06664>. [Online]. Available: <https://doi.org/10.48550/arXiv.1711.06664>
- [7] T. Hubrigtsen, D. Wierichs, E. Gil-Fuster, P.-J. H. S. Derks, P. K. Faehrmann, and J. J. Meyer, "Training quantum embedding kernels on near-term quantum computers," *Physical Review A*, vol. 106, p. 042431, 2022, <https://doi.org/10.1103/PhysRevA.106.042431>. [Online]. Available: <https://doi.org/10.1103/PhysRevA.106.042431>
- [8] G. Gentinetta, A. Thomsen, D. Sutter, and S. Woerner, "The complexity of quantum support vector machines," *Quantum*, vol. 8, p. 1225, 2024, <https://doi.org/10.22331/q-2024-01-11-1225>. [Online]. Available: <https://doi.org/10.22331/q-2024-01-11-1225>
- [9] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, "Variational quantum algorithms," *Nature Reviews Physics*, vol. 3, pp. 625–644, 2021, <https://doi.org/10.1038/s42254-021-00348-9>. [Online]. Available: <https://doi.org/10.1038/s42254-021-00348-9>
- [10] J. M. V. Sindayigaya, "FT-FM: A financial-transaction-specific quantum feature map for variational quantum classifiers in fraud detection," *International Journal of Electronics and Telecommunications*, vol. 72, no. 3, pp. 1–8, 2026, <https://doi.org/10.24425/ijet.2026.1729>. [Online]. Available: <https://doi.org/10.24425/ijet.2026.1729>
- [11] —, "FT-FM-lite: Tunable-entanglement refinement and empirical benchmarking on financial-cybersecurity datasets," 2026, manuscript submitted for publication; no DOI assigned.
- [12] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Foundations and Trends in Machine Learning*, vol. 16, no. 4, pp. 494–591, 2023, <https://doi.org/10.1561/2200000101>. [Online]. Available: <https://doi.org/10.1561/2200000101>
- [13] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995, <https://doi.org/10.1007/BF00994018>. [Online]. Available: <https://doi.org/10.1007/BF00994018>
- [14] M. Schuld, "Supervised quantum machine learning models are kernel methods," 2021, <https://doi.org/10.48550/arXiv.2101.11020>. [Online]. Available: <https://doi.org/10.48550/arXiv.2101.11020>

- [15] S. Thanasilp, S. Wang, M. Cerezo, and Z. Holmes, “Exponential concentration in quantum kernel methods,” *Nature Communications*, vol. 15, p. 5200, 2024, <https://doi.org/10.1038/s41467-024-49287-w>. [Online]. Available: <https://doi.org/10.1038/s41467-024-49287-w>
- [16] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, “Connecting ansatz expressibility to gradient magnitudes and barren plateaus,” *PRX Quantum*, vol. 3, no. 1, p. 010313, 2022, <https://doi.org/10.1103/PRXQuantum.3.010313>. [Online]. Available: <https://doi.org/10.1103/PRXQuantum.3.010313>
- [17] R. Shaydulin and S. M. Wild, “Importance of kernel bandwidth in quantum machine learning,” *Physical Review A*, vol. 106, no. 4, p. 042407, 2022, <https://doi.org/10.1103/PhysRevA.106.042407>. [Online]. Available: <https://doi.org/10.1103/PhysRevA.106.042407>
- [18] N. Liu and P. Rebentrost, “Quantum machine learning for quantum anomaly detection,” *Physical Review A*, vol. 97, p. 042315, 2018, <https://doi.org/10.1103/PhysRevA.97.042315>. [Online]. Available: <https://doi.org/10.1103/PhysRevA.97.042315>
- [19] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, “Calibrating probability with undersampling for unbalanced classification,” in *Proc. IEEE Symposium Series on Computational Intelligence*, 2015, pp. 159–166, <https://doi.org/10.1109/SSCI.2015.33>. [Online]. Available: <https://doi.org/10.1109/SSCI.2015.33>
- [20] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, <https://doi.org/10.1023/A:1010933404324>. [Online]. Available: <https://doi.org/10.1023/A:1010933404324>
- [21] Qiskit contributors, “Qiskit: An open-source framework for quantum computing,” 2023, <https://doi.org/10.5281/zenodo.2573505>. [Online]. Available: <https://doi.org/10.5281/zenodo.2573505>
- [22] F. Wilcoxon, “Individual comparisons by ranking methods,” *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945, <https://doi.org/10.2307/3001968>. [Online]. Available: <https://doi.org/10.2307/3001968>
- [23] S. Holm, “A simple sequentially rejective multiple test procedure,” *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979, <https://doi.org/10.2307/4615733>. [Online]. Available: <https://doi.org/10.2307/4615733>