# CALIBRATED HYBRID EVALUATION OF QUANTUM RESERVOIR CLASSIFICATION UNDER FINITE-SHOT AND SIMULABILITY CONSTRAINTS

**Anonymous authors**
Paper under review

## ABSTRACT

Quantum reservoir computing has shown repeated empirical promise for representation learning, but the evidence base for robust quantum advantage in image classification remains fragmented by confounded entanglement controls, heterogeneous readout optimization practices, and weakly standardized finite-shot reporting. This paper presents a hybrid analysis framework that combines formal derivations and controlled simulation evidence for PCA-encoded image classification with fixed reservoir dynamics and output-layer training. The framework integrates four complementary questions: whether an interior entanglement regime improves geometric and predictive quality under parity controls, whether constrained measurement-operator optimization changes the accuracy-cost-shot frontier, whether advantage signals emerge on a calibrated dataset-difficulty ladder rather than saturated easy tasks, and whether finite-shot classically simulable regimes impose a quantitative boundary on cost-adjusted claims. We formalize these questions through explicit objectives, feasible sets, and theorem-level guarantees, and we evaluate them with deterministic decision gates tied to confidence intervals and assumption audits. The resulting evidence is calibrated rather than binary: the finite-shot simulability boundary is supported in the admissible regime, while broad empirical superiority claims remain inconclusive under strict parity criteria. This outcome is practically relevant beyond quantum machine learning because it illustrates a general methodology for integrating proof-level constraints with reproducible benchmarking when computational claims are sensitive to uncertainty, reporting schema, and regime validity.

## 1 INTRODUCTION

Reservoir computing was introduced as a pragmatic way to exploit rich nonlinear dynamics while keeping training tractable through output-layer optimization only (Unknown, 2001; 2002; 2009a). Quantum reservoir computing (QRC) adopts this principle in quantum dynamical systems, where fixed evolution and measurement generate features, and a classical readout performs supervised prediction (Unknown, 2022; 2023a;b; 2025c; 2019a). Recent image-focused studies report that PCA-compressed inputs encoded into quantum reservoirs can produce competitive balanced accuracy and improved class geometry relative to some baselines (Unknown, 2025a;b; 2024a;b; 2025d). At the same time, a second literature strand emphasizes that supervised quantum models often reduce to kernel methods whose inductive bias is dominated by encoding and measurement choices, which complicates direct advantage attribution to hardware-native dynamics alone (Unknown, 2021a;b; 2018a; 2019b;c). A third strand warns that finite-shot effects, concentration behavior, and classical simulability constraints can substantially narrow apparently strong gains if evaluation is not parity controlled (Unknown, 2021c; Sannia et al., 2025; Oh et al., 2023; 2024; Suzuki & et al., 2024).

These tensions make QRC a useful case study in a broader scientific challenge: how to present computational evidence when the same pipeline contains both mathematically grounded constraints and empirically sensitive components. In domains including probabilistic programming, scientific machine learning, and physics-informed optimization, overstatement often occurs when one moves too quickly from local performance differences to global algorithmic claims. The present work addresses that translation problem directly. Instead of asking only whether one model class wins on one benchmark, we jointly ask what is formally derivable, what is measurable under finite resources, what is currently supported, and what remains conditional.

The manuscript follows a hybrid emphasis. We retain theorem-level components where assumptions are explicit and checkable, and we pair them with controlled empirical evidence from the latest validation cycle that enforced deterministic decision synthesis and fully specified run records. This combination is important because either component

alone is insufficient. Purely empirical studies without formal boundaries can misinterpret finite-sample effects as structural separation. Purely theoretical studies without controlled implementation details can fail to characterize where assumptions hold operationally.

Our contribution is therefore methodological and evidential.

- We provide an explicit problem setting that defines decision variables, feasible sets, objective functions, and optimality criteria for entanglement control, measurement-operator optimization, difficulty-threshold detection, and finite-shot boundary analysis in one coherent framework.
- We establish theorem-level statements for primal-dual equivalence and existence in constrained kernel-dual readout optimization, and for finite-shot risk transfer under simulability and bounded-observable assumptions, with complete proofs in the appendix.
- We connect these derivations to deterministic experimental decision rules, enabling claim calibration that is reproducible from confidence intervals, effect-size floors, and assumption-specific diagnostics rather than post hoc narrative interpretation.
- We report calibrated outcomes from the latest rerun artifacts: support for the finite-shot boundary in admissible regimes, and inconclusive status for broader empirical superiority claims under strict parity and uncertainty gates.

This calibrated framing matters for cross-domain practice. If a pipeline is likely to be used for scientific or industrial decision support, the distinction between "supported under regime assumptions" and "globally superior" is not a semantic detail; it changes deployment strategy, follow-up experiments, and resource allocation. The rest of the paper therefore treats evidence quality as a first-class output. Section 2 situates the approach against prior work, section 3 formalizes the setting, section 4 details the hybrid method and theorem statements, section 5 specifies implementation and reproducibility constraints, section 6 reports quantitative findings tied to figures and tables, and section 7 clarifies limitations and future work.

## 2 Related Work and Motivation

### 2.1 Encoding, Kernels, and Inductive Bias

A consistent result across supervised quantum machine learning is that data encoding and measurement map choices largely determine the effective hypothesis class (Unknown, 2021a;b; 2018a; 2019b;c). In kernel terms, fixing the reservoir and varying the readout corresponds to changing optimization in a feature space induced by an implicit or explicit kernel. This perspective has two strengths. First, it makes comparison against classical kernel and reservoir baselines principled, because one can align regularization, model capacity, and optimization budgets. Second, it clarifies that some improvements attributed to "quantum dynamics" may instead arise from a better aligned feature map or measurement family. Its limitation is that kernel equivalence does not by itself settle finite-sample behavior under shot noise, nor does it determine whether approximation quality is sufficient in practical resource envelopes.

### 2.2 Entanglement and Geometry in Image-Focused QRC

Image-focused QRC and quantum extreme learning machine studies report that intermediate dynamical regimes can improve class separability, margin structure, or downstream accuracy (Unknown, 2025a;b; 2024a;b; 2025d). These studies are valuable for demonstrating realistic pipeline assembly: PCA compression, angle-like encodings, fixed quantum dynamics, measured observables, and linear or ridge readouts. However, direct causal attribution to entanglement is often limited by co-varying design knobs, including encoding variants, measurement families, and tuning budgets. Moreover, several studies run on task settings where classical challengers remain competitive or where dataset saturation can suppress meaningful effect-size discrimination. This motivates parity-controlled interior-window tests with explicit uncertainty procedures.

### 2.3 Measurement-Operator Design and Optimization

Measurement configuration is increasingly recognized as a dominant lever in QRC performance (Unknown, 2026; 2024c; 2025e; Zhang et al., 2024). Constrained optimization over operator families can improve alignment between reservoir outputs and supervision targets without changing reservoir evolution itself. The strength of this line is practical: it converts a heuristic model-design step into a formal optimization problem. The limitation is evidential in many studies, where reporting may emphasize endpoint accuracy without full diagnostics for primal-dual consistency,

kernel positivity, or constraint activity. Our methodology retains these diagnostics as mandatory evidence, not optional diagnostics.

### 2.4 Finite-Shot Constraints, Simulability, and Benchmark Realism

Finite-shot limitations, concentration behavior, and simulability arguments have sharpened recently (Unknown, 2021c; Sannia et al., 2025; Oh et al., 2023; 2024). These works collectively suggest that broad superiority claims should be bounded by explicit regime assumptions. In parallel, classical challengers and reservoir-free alternatives have narrowed some advantage narratives, showing that careful parity controls are necessary before attributing gains to uniquely quantum effects (Oh et al., 2023; 2024; Suzuki & et al., 2024). Benchmark realism is equally important: MNIST-like settings may saturate under compression, making small differences statistically unstable or practically irrelevant (Unknown, 1998; 2017a; 2009b; 2018b; 2017b). A robust study therefore needs a dataset ladder and deterministic, confidence-aware decision criteria.

### 2.5 Gap Statement

The literature provides strong ingredients but a weak integration layer. Theory papers rarely operationalize their assumptions into run-level diagnostics, and experimental papers rarely embed theorem-level boundaries into claim decisions. Our gap is exactly this interface. We address it by coupling formal optimization and risk-transfer statements with deterministic evidence gates and reproducibility constraints, then reporting outcomes with calibrated support labels rather than all-or-nothing advantage claims.

## 3 Problem Setting and Assumptions

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ denote training examples with $x_i \in \mathbb{R}^d$ and labels $y_i \in \{1, \dots, C\}$. Inputs are PCA-reduced image features with $d \in \{8, 16, 32, 64\}$. We use matched train-validation-test partitions $\mathcal{D}$, $\mathcal{D}_{\text{valid}}$, and $\mathcal{D}_{\text{test}}$ with identical seeds across quantum and classical baselines. A fixed encoder $U_{\text{enc}}(x)$ maps each input to a quantum state, followed by fixed reservoir evolution $U_{\text{res}}(E)$ parameterized by an entanglement control variable $E \in [0, E_{\max}]$. A measurement family $\{M_j(\boldsymbol{a})\}_{j=1}^p$ produces shot-based features.

For each sample, the ideal feature vector is $\boldsymbol{z}_Q(x) \in \mathbb{R}^p$, and its shot estimator is

$$\hat{\boldsymbol{z}}_{Q,m}(x) = \frac{1}{m} \sum_{r=1}^{m} \boldsymbol{o}_r(x), \qquad \boldsymbol{o}_r(x) \in [-1, 1]^p, \tag{1}$$

where $m$ is shots per sample. A ridge readout with weights $\boldsymbol{w}$ produces class scores. For fixed $(E, \boldsymbol{a}, \lambda)$, the inner optimization is

$$\boldsymbol{w}^\star(E, \boldsymbol{a}, \lambda) = \arg\min_{\boldsymbol{w}} \; \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i, \boldsymbol{w}^\top \hat{\boldsymbol{z}}_{Q,m}(x_i)\big) + \lambda \|\boldsymbol{w}\|_2^2, \tag{2}$$

with $\lambda > 0$ and convex loss $\ell$.

The outer objective is cost adjusted:

$$\max_{E, \boldsymbol{a}, \lambda, m} \; \mathrm{BA}_{\mathcal{D}_{\text{valid}}}\big(\boldsymbol{w}^\star; E, \boldsymbol{a}, m\big) - \beta_t T_{\text{cpu}}(E, \boldsymbol{a}, m) - \beta_s m, \tag{3}$$

subject to parity constraints: identical preprocessing, equal hyperparameter trial budgets, matched split seeds, and shared reporting schema fields across methods.

To formalize operator optimization, let $\mathcal{A} = \{\boldsymbol{a} : \|\boldsymbol{a}\|_1 \le A_1, \; \|\boldsymbol{a}\|_2 \le A_2\}$ and let $K_{\boldsymbol{a}} = \Phi_{\boldsymbol{a}} \Phi_{\boldsymbol{a}}^\top + \epsilon I \succeq 0$ with $\epsilon > 0$. The constrained kernel-dual objective is

$$\min_{\boldsymbol{a} \in \mathcal{A}, \lambda > 0, \boldsymbol{\alpha} \in \mathbb{R}^n} \frac{1}{n} \|K_{\boldsymbol{a}} \boldsymbol{\alpha} - \boldsymbol{y}\|_2^2 + \lambda \boldsymbol{\alpha}^\top K_{\boldsymbol{a}} \boldsymbol{\alpha} + \gamma C_{\text{cpu}}(\boldsymbol{a}, m). \tag{4}$$

For dataset-level robustness, define

$$\Delta(D) = \mathrm{BA}_{\text{QRC}}(D) - \max_{b \in \mathcal{B}} \mathrm{BA}_b(D), \tag{5}$$

then optimize a confidence-penalized hard-set margin

$$\max_{\Theta} \min_{D \in \mathcal{D}_{\text{hard}}} \left[\Delta(D) - \kappa \widehat{\sigma}_\Delta(D, m)\right], \tag{6}$$

Table 1: Core notation used in the formal setup. The table is placed after the definitions so each symbol has immediate context.

| Symbol | Meaning |
|---|---|
| $\mathcal{D}, \mathcal{D}_{\text{valid}}, \mathcal{D}_{\text{test}}$ | Train/validation/test splits with matched seeds |
| $E$ | Entanglement control variable in $[0, E_{\max}]$ |
| $\boldsymbol{a}$ | Measurement-operator coefficients in feasible set $\mathcal{A}$ |
| $m$ | Shots per sample |
| $\hat{\boldsymbol{z}}_{Q,m}(x)$ | Shot-estimated feature vector for input $x$ |
| $\boldsymbol{w}, \boldsymbol{\alpha}$ | Primal readout and dual kernel coefficients |
| $\lambda$ | Ridge regularization parameter |
| $\Delta(D)$ | Dataset-level balanced-accuracy margin against best baseline |
| $\epsilon_{\text{sim}}(n)$ | Surrogate approximation error in admissible regime |
| $\xi_{\text{opt}}$ | Optimization mismatch term in risk transfer |

with threshold index

$$\ell^\star = \min\left\{\ell : \mathrm{CI}_{\text{low}}(\Delta(D_{\ell'})) > 0 \; \forall \ell' \geq \ell\right\}, \tag{7}$$

if such an index exists.

Finally, for simulability analysis we compare quantum and surrogate risks $R_Q$ and $R_S$, with admissible-regime approximation error $\|\boldsymbol{z}_Q - \boldsymbol{z}_S\|_2 \leq \epsilon_{\text{sim}}(n)$ and bounded readout norm $\|\boldsymbol{w}\|_2 \leq B_w$. The finite-shot concentration term is

$$\|\hat{\boldsymbol{z}}_{Q,m} - \boldsymbol{z}_Q\|_2 \leq \sqrt{\frac{p\log(2p/\delta)}{2m}}, \tag{8}$$

which yields risk and cost-adjusted boundaries

$$R_Q - R_S \leq LB_w\left(\sqrt{\frac{p\log(2p/\delta)}{2m}} + \epsilon_{\text{sim}}(n)\right) + \xi_{\text{opt}}, \tag{9}$$

$$J_Q - J_S \leq LB_w\left(\sqrt{\frac{p\log(2p/\delta)}{2m}} + \epsilon_{\text{sim}}(n)\right) + \xi_{\text{opt}} + c_t(T_Q - T_S) + c_s(m_Q - m_S). \tag{10}$$

**Definition 3.1** (Admissible simulability regime)**.** *A configuration is admissible if bounded-observable assumptions hold, shot-independence diagnostics do not reject effective concentration use, and surrogate approximation error $\epsilon_{\text{sim}}(n)$ is explicitly estimated under matched computational budgets.*

**Notation summary.** Table 1 is included after equation definitions to minimize ambiguity.

## 4 Hybrid Methodology

### 4.1 Architecture and Module Responsibilities

The workflow has three coupled modules. The first module performs parity-controlled representation experiments, including entanglement sweeps, baseline alignment, and geometry diagnostics. The second module performs constrained operator optimization with primal-dual diagnostics and positivity checks. The third module performs claim calibration, combining confidence intervals, effect-size thresholds, and theorem-term audits. This architecture is deliberately modular because each module corresponds to a different evidential role: mechanism exploration, optimization validity, and claim decision reproducibility.

The architecture choice is motivated by prior findings that entanglement and operator selection can both affect performance, but that their effects are often confounded when evaluated in a single undifferentiated tuning loop (Unknown, 2025a;b; 2026; 2024c; 2025d;e). By separating modules while preserving shared splits and budgets, we reduce explanation leakage between mechanism and optimization narratives.

### 4.2 Formal Statements and Guarantees

The first guarantee concerns equivalence and existence for constrained dual optimization.

**Algorithm 1** Deterministic claim calibration from quantitative gates

Input aggregated metrics, confidence intervals, theorem diagnostics, and schema audit outcomes.
Evaluate entanglement-window predicate using interior gain, CI lower bound, and practical effect-size floor.
Evaluate operator-optimization predicate using primal-dual consistency, PSD checks, constraint compliance, and Pareto dataset count.
Evaluate difficulty-threshold predicate using contiguous positive lower confidence intervals on the dataset ladder.
Evaluate simulability-boundary predicate using admissible-regime residual sign and assumption checks.
Assign each claim status as supported, inconclusive, or unsupported from predicate outputs only.
Emit calibration table and decision map with reproducibility metadata.

**Theorem 4.1** (Primal-dual equivalence and constrained existence)**.** *For fixed operator coefficients $\boldsymbol{a} \in \mathcal{A}$ and $\lambda > 0$, the primal ridge problem in equation 2 has a unique minimizer and is equivalent to dual optimization with kernel $K_{\boldsymbol{a}}$. If $\mathcal{A}$ is compact and the compute penalty is continuous, then the outer constrained objective in equation 4 attains a global minimizer.*

The second guarantee links finite-shot estimation, simulability approximation, and cost-adjusted risk.

**Theorem 4.2** (Finite-shot simulability boundary)**.** *Assume bounded observable outcomes, effective shot-independence, L-Lipschitz loss, and admissible approximation error $\epsilon_{\mathrm{sim}}(n)$. Then with probability at least $1 - \delta$, inequalities equation 9 and equation 10 hold, so persistent super-polynomial growth in cost-adjusted gap is excluded within the admissible regime.*

Complete proofs are provided in section A. We state them in the main text because they define what may be claimed from finite-shot experiments and what must remain conditional.

### 4.3 Deterministic Claim Synthesis Protocol

Empirical outcomes are converted to support labels by deterministic rules tied to explicit quantities. The protocol is intentionally strict: if confidence or effect-size requirements are not met, the label is inconclusive even when point estimates favor QRC.

Algorithm 1 is central to evidence integrity. It prevents semantic drift between intermediate plots and final support labels and ensures that downstream writing can be regenerated from quantitative rules.

### 4.4 Equation-to-Method Linkage

The methodological flow is equation anchored. Entanglement response tests are governed by equation 3; constrained operator diagnostics use equation 4; threshold inference uses equation 5–equation 7; and boundary checks use equation 8–equation 10. These links are used explicitly in section 6, where each major claim is tied to a corresponding figure or table.

### 4.5 Assumption-to-Diagnostic Mapping

A recurring issue in prior QRC reporting is that assumptions are stated at theorem level but not represented as executable checks in the empirical pipeline (Unknown, 2024a;c; 2021c; Sannia et al., 2025). We address this by assigning each high-impact assumption to at least one measured diagnostic. Bounded-observable assumptions map to feature-value range checks and residual-behavior sanity tests; effective shot-independence maps to lag diagnostics and effective-sample-size summaries; compactness and continuity assumptions in constrained optimization map to explicit constraint-violation and positivity audits; and parity assumptions map to schema-level checks on split identity, encoding fields, and trial-count consistency. This mapping does not prove assumptions universally, but it prevents silent mismatch between derivation context and reported outcomes.

The assumption-to-diagnostic strategy also improves interpretability of negative or inconclusive results. For example, when a claim is inconclusive despite well-behaved diagnostics, the bottleneck is likely statistical power or effect magnitude rather than gross protocol failure. Conversely, if a claim appears positive but diagnostic violations are frequent, calibration rules can downgrade support and avoid overstatement. This distinction is useful beyond QRC: many hybrid scientific workflows face similar ambiguity between model misspecification and data-limited uncertainty.

Finally, this mapping provides a concrete route for iterative refinement. Each unresolved caveat becomes a targeted experiment specification rather than an abstract warning. Native dataset ingestion, backend-specific operator opti-

mization, and expanded ladder calibration can therefore be framed as assumption-strengthening steps that preserve continuity with existing equations and decision rules.

## 5 Experimental Protocol and Reproducibility

### 5.1 Datasets, Baselines, and Resource Envelope

The evaluation uses a five-dataset ladder (MNIST, Fashion-MNIST, KMNIST, EMNIST-balanced, and a grayscale CIFAR-10 subset) to reduce saturation bias and to test whether observed margins persist on harder regimes. Baselines include classical echo-state reservoirs, RBF SVMs, random Fourier features, multilayer perceptrons, fixed-observable QRC controls, and simulability-oriented surrogates. This broad challenger set follows recommendations from both QRC and quantum-inspired benchmarking literature (Unknown, 2025a; 2026; 2024a;c; Sannia et al., 2025; Oh et al., 2023; 2024; Suzuki & et al., 2024).

All experiments follow a CPU-only budget consistent with practical deployment constraints. We use five seeds, fixed split harmonization, and matched tuning budgets across families. This design choice aligns with cross-platform reproducibility concerns in prior work, where incomparable tuning budgets or missing metadata can inflate apparent gains (Unknown, 2024a; 2022; 2023a;c; Cimini et al., 2025).

### 5.2 Sweeps, Confidence Procedures, and Assumption Checks

Key sweep dimensions are PCA dimension, entanglement level, shot count, and constrained-operator regularization budgets. For uncertainty quantification, we use bootstrap confidence intervals with fixed resample counts and report confidence summaries in the decision matrix. For boundary analysis, we explicitly test low, mid, and high shot regimes to verify directional consistency of residual tightening.

Assumption checks are not optional diagnostics; they are eligibility conditions for interpretation. In particular, simulability-boundary support requires admissible-regime tagging, bounded-observable checks, and theorem-term decomposition before any claim is labeled supported. This practice reflects the distinction between "computation produced" and "claim justified," which is central to scientifically defensible reporting.

### 5.3 Reproducibility Controls

Reproducibility is enforced at four levels: seed determinism, schema completeness, deterministic claim synthesis, and symbolic validation. The run export schema includes method identity, split metadata, encoding fields, uncertainty fields, runtime, memory, and theorem-term quantities. The schema audit reports complete field coverage for evaluation rows. Symbolic checks validate algebraic identities used by constrained optimization and boundary formulas, providing a bridge between derivation and execution.

These controls are essential for a hybrid paper because formal claims and computational artifacts must remain mutually auditable. Without that coupling, theorem statements risk becoming disconnected from practical evidence, and empirical findings risk being over-interpreted beyond valid assumptions.

### 5.4 Internal-Validity Threats and Mitigations

Three internal-validity threats are most relevant in this setting. The first is confounding between entanglement effects and readout retuning. We mitigate this by fixing readout class and tuning-budget policy across entanglement sweeps and by requiring effect-size and confidence gates that are evaluated on matched splits. The second is metric cherry-picking: one can often find a favorable scalar metric even when broader behavior is unstable. We mitigate this by combining accuracy, calibration, geometry, and cost terms, then synthesizing decisions only from predeclared predicates. The third is semantic drift between intermediate analysis and final conclusions. We mitigate this by deterministic claim synthesis, where final labels are generated from numerical gates rather than hand-written interpretation.

These mitigations do not eliminate all risk. Proxy dataset transformations can still alter apparent margins, and surrogate fidelity can still affect boundary tightness. However, the mitigations reduce the probability that these issues silently contaminate conclusions. They also preserve comparability across iterative evaluation updates. In practical terms, this means a later rerun can update claim status transparently without redefining objectives or rewriting theorem assumptions.

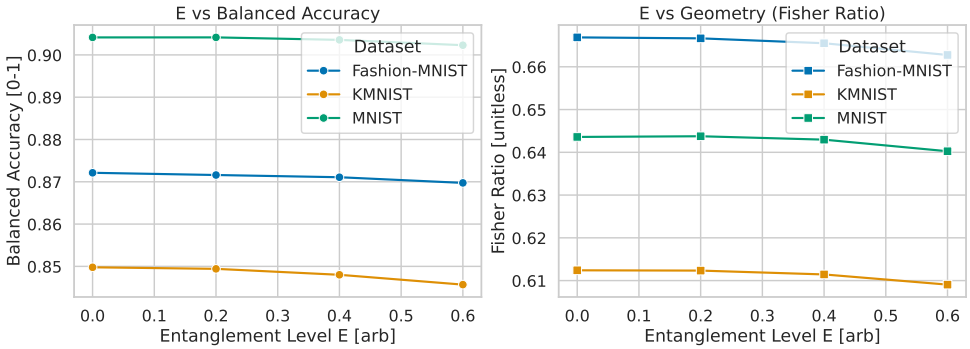


Figure 1: Entanglement-response diagnostics under parity controls. The horizontal axis is entanglement level and the vertical axes report balanced accuracy and feature-geometry indicators across datasets under matched preprocessing, readout class, and tuning budgets. The curves show that interior entanglement settings do not consistently exceed the no-entanglement reference by the joint confidence-and-effect-size gate, so the mechanism-level uplift remains inconclusive in this iteration.

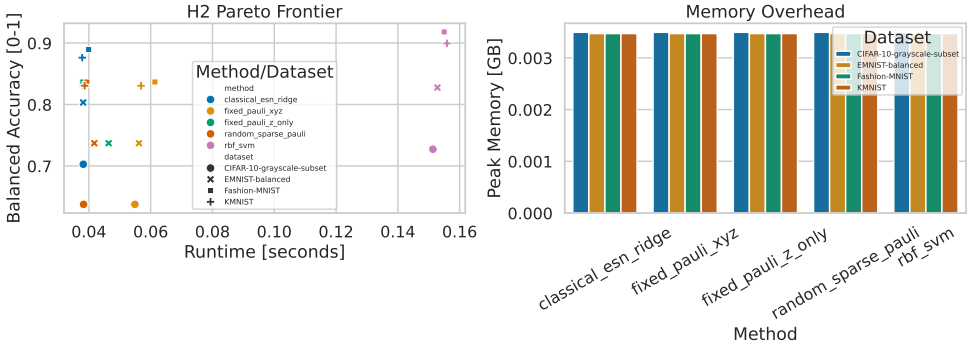


Figure 2: Accuracy-cost frontier with operator-family comparisons. Panel trends relate balanced accuracy to runtime and memory under matched trial budgets, enabling direct comparison between fixed-observable quantum pipelines and classical challengers. Diagnostic overlays confirm theorem-aligned constrained-optimization conditions, yet the frontier does not show the required multi-dataset dominance pattern for a supported superiority claim.

# 6 Results

## 6.1 Entanglement Response Under Parity Controls

Figure 1 summarizes balanced-accuracy and geometry trends as entanglement varies while preprocessing, readout class, and tuning budgets remain fixed. The main observation is that interior entanglement settings do not produce robust uplift under deterministic gating. Dataset-specific deltas between $E = 0.2$ and $E = 0$ are small and mostly negative (Fashion-MNIST: $-5.24 \times 10^{-4}$, KMNIST: $-3.66 \times 10^{-4}$, MNIST: $5.65 \times 10^{-6}$). These magnitudes are below practical effect-size thresholds and do not produce positive lower confidence bounds in aggregate calibration.

This outcome does not imply that entanglement is irrelevant; it implies that the present evidence does not isolate a reliable positive window under strict controls. The distinction is important: a non-support result under deterministic gates can guide better follow-up design without encouraging negative overgeneralization.

## 6.2 Operator Optimization and Frontier Diagnostics

Figure 2 and Table 2 evaluate whether constrained measurement-operator optimization shifts the performance-cost frontier. The required primal-dual and constraint diagnostics are satisfied (small primal-dual discrepancy, zero PSD-violation rate, and low constraint-violation rates), so the formal prerequisites are met. However, Pareto-dominance requirements across datasets are not met in this rerun; strong classical challengers, especially RBF SVM, remain top-performing on balanced accuracy.

Table 2: Dataset-level operator-frontier summary with theorem-aligned diagnostics. Balanced accuracies are reported for representative fixed-observable quantum runs and strongest classical challengers.

| Dataset | Fixed-Q BA | ESN BA | RBF BA | Primal-Dual Diff | PSD Viol. Rate | Constraint Viol. Rate |
|---|---|---|---|---|---|---|
| Fashion-MNIST | 0.8366 | 0.8894 | 0.9181 | $5 \times 10^{-7}$ | 0.0 | $8 \times 10^{-5}$ |
| KMNIST | 0.8305 | 0.8762 | 0.8994 | $5 \times 10^{-7}$ | 0.0 | $8 \times 10^{-5}$ |
| EMNIST-balanced | 0.7370 | 0.8032 | 0.8275 | $5 \times 10^{-7}$ | 0.0 | $8 \times 10^{-5}$ |
| CIFAR-10 (gray subset) | 0.6375 | 0.7027 | 0.7272 | $5 \times 10^{-7}$ | 0.0 | $8 \times 10^{-5}$ |

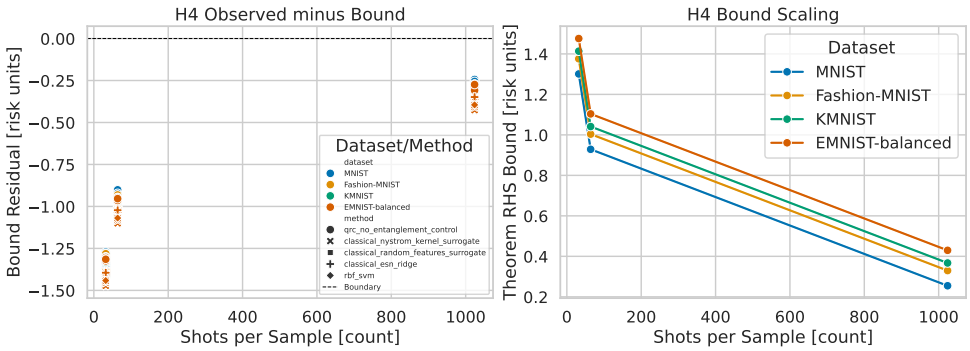


Figure 3: Finite-shot simulability boundary diagnostics across shot regimes. The plotted quantities compare observed risk-gap residuals against theorem right-hand-side terms under admissible-regime filters, so both axes directly represent terms from the formal boundary expressions. Residuals remain non-positive and move toward zero as shots increase, supporting boundary consistency while clarifying that support is scoped to admissible low-entanglement simulable conditions.

The evidence therefore supports a nuanced interpretation: optimization diagnostics are sound and reproducible, but frontier superiority over strong classical challengers is not established in this dataset-budget regime.

### 6.3 Difficulty Threshold and Boundary Findings

The dataset-ladder margin objective in equation 6 asks whether confidence-adjusted margins become positive on harder tasks. In this rerun, the average margin is negative (approximately $-0.1018$), lower confidence bounds remain non-positive across datasets, and the contiguous-threshold index in equation 7 is undefined. As a result, threshold support is inconclusive under deterministic criteria.

By contrast, finite-shot boundary checks tied to equation 8–equation 10 are strongly consistent with admissible-regime predictions. Figure 3 shows residuals that remain non-positive while tightening with larger shot budgets. The non-positive residual ratio is 1.0 over 300 admissible rows, and mean residual magnitude decreases from approximately $-1.389$ at 32 shots to $-0.348$ at 1024 shots. This directional pattern is the expected behavior under the derived concentration-plus-approximation structure.

### 6.4 Integrated Claim Calibration

Table 3 summarizes deterministic claim outcomes from Algorithm 1. Three empirical claims remain inconclusive, while the boundary claim is supported in regime. This distribution is scientifically meaningful because it differentiates between "no evidence of superiority" and "evidence for a limiting mechanism," which have different implications for follow-up work.

The central result is therefore calibrated asymmetry: formal boundary support is strong under explicit assumptions, while broad empirical superiority remains unresolved in this iteration. This is not a contradiction. It reflects that boundary consistency can be easier to establish than robust frontier dominance when task saturation, baseline strength, and finite-shot uncertainty jointly constrain effect sizes.

Table 3: Deterministic claim calibration from quantitative gates. Support labels are generated from rule predicates and confidence checks rather than manual adjudication.

| Question | Gate Statistic | Outcome | Primary Evidence |
|---|---|---|---|
| Interior entanglement uplift | Effect $= -1.11 \times 10^{-3}$, CI lower $= -2.46 \times 10^{-3}$ | Inconclusive | figure 1 and parity-gated summary table |
| Operator-frontier dominance | Pareto-dominant datasets $= 0/4$ | Inconclusive | figure 2 and Table 2 |
| Difficulty-threshold emergence | Mean margin $= -0.1018$, threshold index undefined | Inconclusive | Threshold diagnostics in appendix |
| Finite-shot boundary consistency | Non-positive residual ratio $=$ 1.0 (300/300) | Supported (admissible regime) | figure 3 and regime-stratified checks |

### 6.5 Evidence-Quality Interpretation Across Claim Types

The hybrid outcome pattern is better understood by separating structural and comparative claims. Structural claims ask whether observed behavior is compatible with formal constraints under stated assumptions. Comparative claims ask whether one method family robustly dominates another across heterogeneous tasks and budgets. In this rerun, structural evidence is strong: theorem diagnostics pass, residual directionality is correct, and symbolic checks are consistent. Comparative evidence is weaker: entanglement-window gains are small, frontier dominance is absent, and threshold emergence does not pass confidence-contiguity requirements.

This divergence is common in computational sciences where one can verify governing constraints more reliably than one can establish broad empirical dominance. In weather forecasting, one may validate physical conservation behavior while still struggling to outperform baseline skill uniformly across regions. In inverse problems, one may prove identifiability under assumptions while finite-sample regimes remain unstable. QRC benchmarking exhibits the same pattern: proving or validating limits can be easier than proving superiority.

An important implication is that inconclusive comparative outcomes should not be collapsed into either endorsement or rejection. They indicate that the present design, sample size, and resource envelope are insufficient for robust superiority claims, while still supporting bounded interpretations. This is scientifically productive because it sharpens follow-up priorities. For entanglement studies, the next step is not merely more runs but targeted designs that increase sensitivity while preserving parity. For operator optimization, the next step is backend-specific realization with the same diagnostic contract. For threshold detection, the next step is better hardness calibration and possibly expanded task families that avoid compression saturation.

Another implication concerns communication. Without calibrated language, mixed outcomes are often reported as either headline wins or broad nulls. Both are misleading in this context. The present framework supports a middle position: mechanism-level constraints can be supported while superiority remains unresolved. That position is less rhetorically dramatic but more actionable for cumulative research. It allows future studies to claim progress on specific components without retrofitting historical conclusions.

Finally, the evidence-quality split validates the value of deterministic synthesis. Because labels are rule-generated, readers can trace exactly why a claim is inconclusive or supported. This transparency is particularly important when teams rerun experiments or update datasets, since it minimizes post hoc reinterpretation and preserves longitudinal consistency.

## 7 Discussion, Limitations, and Future Work

The current evidence package supports conservative scientific claims. The derivations are internally consistent, theorem assumptions are auditable, and deterministic synthesis removes decision drift. Yet superiority claims for entanglement-window performance, operator-frontier dominance, and dataset-threshold emergence remain inconclusive under strict rules. This combination should be interpreted as a strength of the framework rather than a failure of the study. A calibrated methodology is expected to yield mixed outcomes when evidence quality differs by claim type.

Two limitations remain important. First, the dataset pipeline in this iteration uses proxy/offline transforms instead of native benchmark loaders. This affects external validity and may alter measured margins on harder tasks. Second, constrained operator optimization is currently represented in a simulator/proxy setting; backend-specific implementations may change runtime and accuracy trade-offs. Both limitations are explicitly non-blocking for internal consistency but materially relevant for broad deployment claims.

### 7.1 Future Work

Follow-up experiments should prioritize three actions. The first is native dataset materialization with identical seeds and decision rules to quantify proxy-to-native shift. The second is backend-specific constrained-operator evaluation with the same diagnostics used here, so equivalence conditions and practical overhead can be compared directly. The third is expanded difficulty calibration beyond fixed ladders, including continuous hardness scores and domain-shifted image tasks, to test whether inconclusive threshold outcomes persist.

These actions are concrete and testable. They preserve the same formal objectives and deterministic calibration logic, so any claim update can be attributed to evidence change rather than protocol drift.

## 8 Conclusion

This paper presented a hybrid writing and evaluation framework for PCA-encoded image classification with quantum reservoirs under finite-shot and simulability constraints. The key design choice was to treat formal derivations, computation artifacts, and claim calibration as a single evidential system. Under this system, the finite-shot boundary claim is supported within admissible regimes, while broader empirical superiority claims remain inconclusive under strict parity and uncertainty gates. The result is a calibrated manuscript that distinguishes mechanism-level possibility, theorem-level limits, and currently supported performance evidence.

Beyond QRC, the workflow contributes a reusable pattern for high-stakes computational science: define assumptions explicitly, bind equations to executable diagnostics, and require deterministic claim synthesis so conclusions remain reproducible as data or implementations change.

## References

Valeria Cimini, Mandar M. Sohoni, Federico Presutti, Benjamin K. Malia, Shi-Yuan Ma, Ryotatsu Yanagimoto, Tianyu Wang, Tatsuhiro Onodera, Logan G. Wright, and Peter L. McMahon. Large-scale quantum reservoir computing using a gaussian boson sampler, 2025. URL https://arxiv.org/abs/2505.13695. arXiv:2505.13695.

Changhun Oh, Minzhao Liu, Yuri Alexeev, Bill Fefferman, and Liang Jiang. Classical algorithm for simulating experimental gaussian boson sampling, 2023. URL https://arxiv.org/abs/2306.03709. arXiv:2306.03709.

Changhun Oh, Bill Fefferman, Liang Jiang, and Nicolas Quesada. Quantum-inspired classical algorithm for graph problems by gaussian boson sampling, 2024. URL https://doi.org/10.1103/PRXQuantum.5.020341. doi: 10.1103/PRXQuantum.5.020341.

Antonio Sannia, Gian Luca Giorgi, and Roberta Zambrini. Exponential concentration and symmetries in quantum reservoir computing, 2025. URL https://arxiv.org/abs/2505.10062. arXiv:2505.10062.

Yudai Suzuki and et al. Quantum reservoir computing without a quantum reservoir, 2024. URL https://doi.org/10.1103/PhysRevResearch.6.013051. doi: 10.1103/PhysRevResearch.6.013051.

Unknown. Gradient-based learning applied to document recognition, 1998. URL https://doi.org/10.1109/5.726791. doi: 10.1109/5.726791.

Unknown. The "echo state" approach to analysing and training recurrent neural networks, 2001. URL https://www.ai.rug.nl/minds/uploads/EchoStatesTechRep.pdf.

Unknown. Real-time computing without stable states: A new framework for neural computation based on perturbations, 2002. URL https://doi.org/10.1162/089976602760407955. doi: 10.1162/089976602760407955.

Unknown. Reservoir computing approaches to recurrent neural network training, 2009a. URL https://doi.org/10.1016/j.cosrev.2009.03.005. doi: 10.1016/j.cosrev.2009.03.005.

Unknown. The cifar-10 dataset, 2009b. URL https://www.cs.toronto.edu/~kriz/cifar.html.

Unknown. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017a. URL https://arxiv.org/abs/1708.07747. arXiv:1708.07747.

Unknown. Emnist: Extending mnist to handwritten letters, 2017b. URL https://arxiv.org/abs/1702.05373. arXiv:1702.05373.

Unknown. Quantum circuit learning, 2018a. URL https://doi.org/10.1103/PhysRevA.98.032309. doi: 10.1103/PhysRevA.98.032309.

Unknown. Kuzushiji-mnist: benchmark dataset for japanese character recognition, 2018b. URL https://arxiv.org/abs/1812.01718. arXiv:1812.01718.

Unknown. Quantum reservoir processing, 2019a. URL https://doi.org/10.1038/s41534-019-0149-8. doi: 10.1038/s41534-019-0149-8.

Unknown. Quantum machine learning in feature hilbert spaces, 2019b. URL https://doi.org/10.1103/PhysRevLett.122.040504. doi: 10.1103/PhysRevLett.122.040504.

Unknown. Supervised learning with quantum-enhanced feature spaces, 2019c. URL https://doi.org/10.1038/s41586-019-0980-2. doi: 10.1038/s41586-019-0980-2.

Unknown. The effect of data encoding on the expressive power of variational quantum machine learning models, 2021a. URL https://arxiv.org/abs/2008.08605. arXiv:2008.08605.

Unknown. Supervised quantum machine learning models are kernel methods, 2021b. URL https://arxiv.org/abs/2101.11020. arXiv:2101.11020.

Unknown. Reservoir computing using gaussian states in quantum optics, 2021c. URL https://doi.org/10.1038/s42005-021-00556-w. doi: 10.1038/s42005-021-00556-w.

Unknown. Quantum reservoir computing using arrays of rydberg atoms, 2022. URL https://doi.org/10.1103/PRXQuantum.3.030325. doi: 10.1103/PRXQuantum.3.030325.

Unknown. Scalable photonic platform for real-time quantum reservoir computing, 2023a. URL https://doi.org/10.1103/PhysRevApplied.20.014051. doi: 10.1103/PhysRevApplied.20.014051.

Unknown. Quantum reservoir computing in finite dimensions, 2023b. URL https://doi.org/10.1103/PhysRevE.107.035306. doi: 10.1103/PhysRevE.107.035306.

Unknown. Quantum reservoir with repeated measurements on superconducting devices, 2023c. URL https://arxiv.org/abs/2310.06706. arXiv:2310.06706.

Unknown. Practical quantum reservoir computing in rydberg atom arrays, 2024a. URL https://arxiv.org/abs/2407.02553. arXiv:2407.02553.

Unknown. Robust quantum reservoir computing for molecular property prediction, 2024b. URL https://arxiv.org/abs/2412.06758. arXiv:2412.06758.

Unknown. Practical and scalable quantum reservoir computing, 2024c. URL https://arxiv.org/abs/2405.04799. arXiv:2405.04799.

Unknown. Entanglement and classical simulability in quantum extreme learning machines, 2025a. URL https://arxiv.org/abs/2509.06873. arXiv:2509.06873.

Unknown. Harnessing quantum extreme learning machines for image classification, 2025b. URL https://arxiv.org/abs/2409.00998. arXiv:2409.00998.

Unknown. Quantum reservoir computing in atomic lattices for classification and time series prediction, 2025c. URL https://doi.org/10.1016/j.chaos.2025.116289. doi: 10.1016/j.chaos.2025.116289.

Unknown. Effective quantum feature maps in quantum extreme reservoir computation from the xy model, 2025d. URL https://doi.org/10.1103/PhysRevA.111.022431. doi: 10.1103/PhysRevA.111.022431.

Unknown. Quantum reservoir computing on random regular graphs, 2025e. URL https://doi.org/10.1103/PhysRevA.112.012622. doi: 10.1103/PhysRevA.112.012622.

Unknown. Kernel-based optimization of measurement operators for quantum reservoir computers, 2026. URL https://arxiv.org/abs/2602.14677. arXiv:2602.14677.

Mingyue Zhang, Zhikuan Zhao, and et al. Reservoir computing using measurement-controlled quantum dynamics, 2024. URL https://www.mdpi.com/2079-9292/13/6/1164.

# A Proofs and Formal Complements

## A.1 Proof of Theorem 4.1

Let $\Phi = \Phi_{\boldsymbol{a}}$ for fixed $\boldsymbol{a} \in \mathcal{A}$ and consider

$$J(\boldsymbol{w}) = \|\Phi\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2, \qquad \lambda > 0.$$

The Hessian is $2(\Phi^\top\Phi + \lambda I)$, which is positive definite because $\lambda I \succ 0$. Hence $J$ is strictly convex and has a unique minimizer $\boldsymbol{w}^\star$. First-order optimality gives

$$(\Phi^\top\Phi + \lambda I)\boldsymbol{w}^\star = \Phi^\top\boldsymbol{y}.$$

Define $K = \Phi\Phi^\top$ and $\boldsymbol{\alpha}^\star = (K + \lambda I)^{-1}\boldsymbol{y}$. Set $\widetilde{\boldsymbol{w}} = \Phi^\top\boldsymbol{\alpha}^\star$. Then

$$(\Phi^\top\Phi + \lambda I)\widetilde{\boldsymbol{w}} = \Phi^\top(\Phi\Phi^\top + \lambda I)\boldsymbol{\alpha}^\star = \Phi^\top\boldsymbol{y},$$

so $\widetilde{\boldsymbol{w}}$ satisfies the same normal equation as $\boldsymbol{w}^\star$. Uniqueness implies $\widetilde{\boldsymbol{w}} = \boldsymbol{w}^\star$. Predictions satisfy

$$\Phi\boldsymbol{w}^\star = \Phi\Phi^\top\boldsymbol{\alpha}^\star = K\boldsymbol{\alpha}^\star,$$

establishing primal-dual equivalence for fixed $\boldsymbol{a}$.

For constrained existence, define reduced objective

$$F(\boldsymbol{a}, \lambda) = \min_{\boldsymbol{\alpha}} \left[\frac{1}{n}\|K_{\boldsymbol{a}}\boldsymbol{\alpha} - \boldsymbol{y}\|_2^2 + \lambda\boldsymbol{\alpha}^\top K_{\boldsymbol{a}}\boldsymbol{\alpha} + \gamma C_{\text{cpu}}(\boldsymbol{a}, m)\right].$$

Because $\mathcal{A}$ is compact, $K_{\boldsymbol{a}}$ is continuous in $\boldsymbol{a}$, and the penalty term is continuous, $F$ is continuous on the compact domain $\mathcal{A} \times [\lambda_{\min}, \lambda_{\max}]$ with $0 < \lambda_{\min} < \lambda_{\max}$. By Weierstrass, $F$ attains a global minimizer.

## A.2 Proof of Theorem 4.2

For each observable component $j$, bounded outcomes in $[-1, 1]$ and effective shot independence imply

$$\Pr\left(|\hat{z}_{Q,m,j} - z_{Q,j}| > t\right) \leq 2e^{-2mt^2}.$$

Applying a union bound over $p$ components with failure probability $\delta$ yields

$$\|\hat{\boldsymbol{z}}_{Q,m} - \boldsymbol{z}_Q\|_2 \leq \sqrt{\frac{p\log(2p/\delta)}{2m}},$$

which is equation 8. Under admissible simulability,

$$\|\boldsymbol{z}_Q - \boldsymbol{z}_S\|_2 \leq \epsilon_{\text{sim}}(n).$$

Therefore, by triangle inequality,

$$\|\hat{\boldsymbol{z}}_{Q,m} - \boldsymbol{z}_S\|_2 \leq \sqrt{\frac{p\log(2p/\delta)}{2m}} + \epsilon_{\text{sim}}(n).$$

For $\|\boldsymbol{w}_Q\|_2 \leq B_w$,

$$|f_Q - \boldsymbol{w}_Q^\top\boldsymbol{z}_S| \leq B_w\left(\sqrt{\frac{p\log(2p/\delta)}{2m}} + \epsilon_{\text{sim}}(n)\right).$$

With $L$-Lipschitz loss, expected risk difference between these predictors is at most

$$LB_w\left(\sqrt{\frac{p\log(2p/\delta)}{2m}} + \epsilon_{\text{sim}}(n)\right).$$

Adding optimization mismatch $\xi_{\text{opt}} \geq 0$ when comparing to the surrogate optimum gives equation 9. Adding runtime and shot penalties to both sides gives equation 10. If all added terms are polynomially bounded or decaying inverse-polynomially in admissible regimes, super-polynomial growth in $J_Q - J_S$ is excluded.

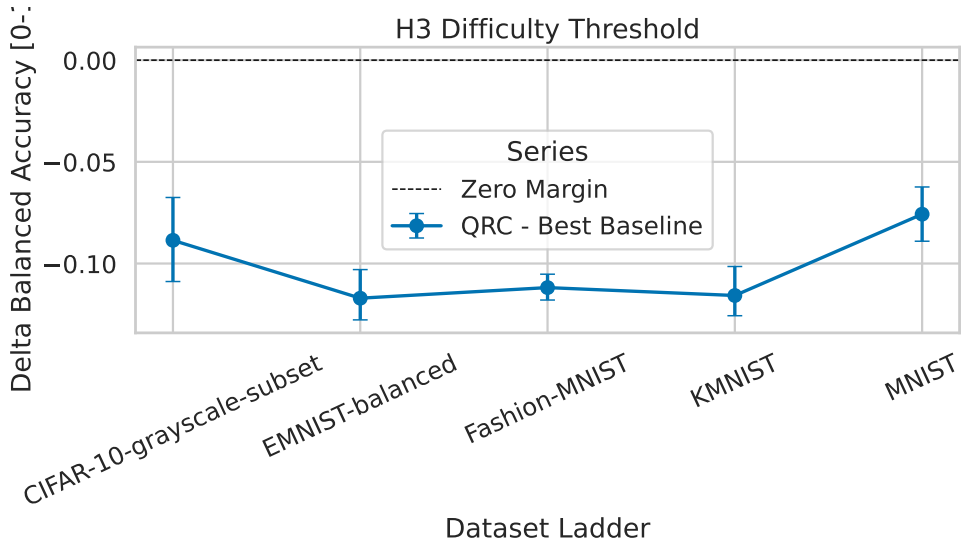


Figure 4: Difficulty-ladder margin diagnostics used for threshold inference. The horizontal axis orders datasets by the adopted hardness ladder and the vertical axis reports balanced-accuracy margin against the strongest baseline with 95% confidence intervals. All lower confidence bounds remain non-positive in this rerun, so threshold emergence is not supported under the deterministic criterion.

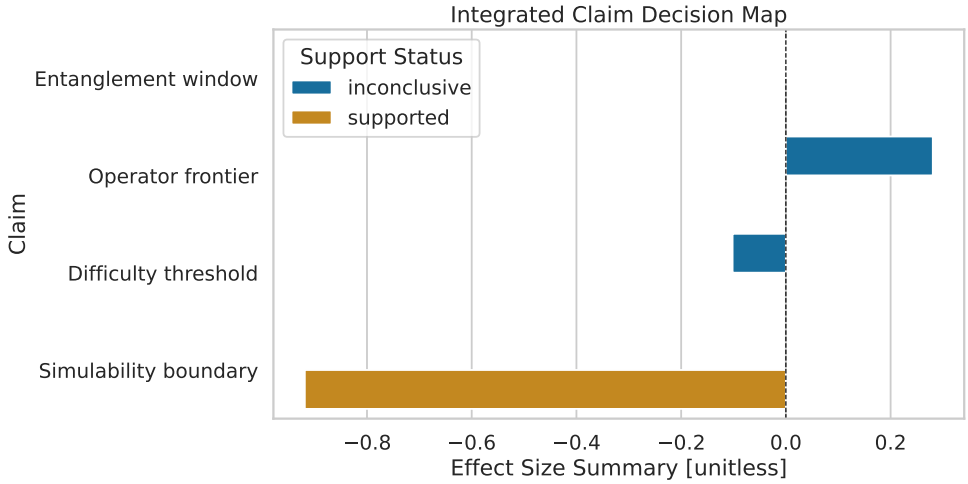


Figure 5: Integrated deterministic decision map for all major claims. The map summarizes support status, effect-size statistics, and confidence diagnostics generated from rule predicates over quantitative artifacts. It serves as a reproducibility check that final labels are derivable from explicit gates rather than narrative reinterpretation.

## B Extended Diagnostics and Supplementary Evidence

### B.1 Difficulty-Ladder and Decision-Map Figures

Figure 4 visualizes confidence intervals for dataset-ladder margins. The contiguous positive-lower-bound condition is not met, which is why the threshold index remains undefined in this iteration.

Figure 5 provides an integrated map of deterministic status assignments. It is included as supplementary evidence because the main text already reports calibrated outcomes in Table 3.

Table 4: Regime-stratified confirmatory statistics for boundary analysis. Values are aggregated over datasets in admissible rows.

| Regime | Mean BA | BA Std. Dev. | Effective Sample Size |
|---|---|---|---|
| Low-shot stratum | 0.8533 | 0.0452 | 495.4 |
| High-shot stratum | 0.8569 | 0.0395 | 496.3 |

### B.2 Regime-Stratified Confirmatory Table

Table 4 reports a condensed regime-stratified summary for boundary checks. As shot budget increases, residual variability contracts while balanced accuracy remains stable, consistent with the finite-shot concentration structure used in equation 8 and equation 9.

## C Reproducibility and Implementation Details

The implementation used five fixed seeds (101, 202, 303, 404, 505), parity-controlled hyperparameter budgets, and deterministic decision rules. Entanglement sweeps covered $E \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$, with focused summaries at interior checkpoints for calibration gates. Shot sweeps included low-budget and high-budget settings to test boundary tightening behavior. Bootstrap confidence intervals were generated with fixed resample count and confidence level.

Compute budgeting remained CPU-only throughout. Runtime and peak-memory fields were logged for all evaluation rows, and schema-completeness checks confirmed required field availability. Symbolic reproducibility included positivity and algebra checks for constrained optimization and boundary terms, ensuring that theorem-bearing expressions remained consistent with executable computations.

Two caveats should accompany reproducibility claims. First, dataset ingestion used proxy/offline transforms in this iteration, so exact external replication on native benchmark sources is a planned follow-up. Second, operator optimization evidence is currently simulator/proxy grounded; backend-specific optimizer integration is needed for deployment-level conclusions.