SIMULABILITY-AWARE QUANTUM RESERVOIR COMPUTING FOR IMAGE CLASSIFICATION UNDER MATCHED READOUT FAIRNESS

Anonymous authors
Paper under review

ABSTRACT

Quantum reservoir computing for vision tasks is often discussed in terms of empirical gains without an equally explicit separation between predictive improvement and computational-advantage claims. We study this issue in a controlled image-classification setting with PCA-compressed inputs, matched preprocessing, and identical linear readout training across classical and quantum reservoirs. The manuscript contributes a hybrid theory-plus-experiment framing: we formalize the readout stage as a strongly convex regularized problem with a unique closed-form optimum under positive regularization, derive concentration-calibrated conditions for reporting positive entanglement regimes, and define a compute-constrained Pareto boundary that filters performance gains by runtime and simulability costs. Experiments on MNIST, Fashion-MNIST, and a grayscale CIFAR-10 variant show that entangling quantum reservoirs can improve average macro-F1 against strong baselines in several configurations, but confidence-qualified positive-regime gates remain unmet under the current proxy-data execution. At the same time, symbolic and numerical checks validate the formal claims, and fixed-universe frontier audits satisfy monotonicity and nondominance requirements. The resulting evidence supports a simulability-safe conclusion: current results are strongest for formal guarantees and boundary/null characterization, while positive advantage claims remain conditional and require canonical-data reruns.

1 INTRODUCTION

Quantum reservoir computing (QRC) is a fixed-dynamics learning paradigm in which quantum evolution provides a nonlinear feature map and only a classical readout is trained. This fixed-reservoir structure is attractive in regimes where gradient-based variational training can suffer optimization pathologies, including barren-plateau effects in highly entangling ansätze (Marrero et al., 2021; Cerezo & et al., 2024). At the same time, modern kernel interpretations of supervised quantum models imply that raw performance differences must be interpreted relative to the induced feature geometry, baseline strength, and data regime (Schuld, 2021; Lloyd et al., 2020; Havlicek et al., 2019; Schuld & Killoran, 2019). For practical image classification, these concerns are amplified by preprocessing decisions: PCA compression can eliminate nuisance variation and may collapse an already easy dataset into a near-ceiling benchmark where any additional representational complexity provides limited measurable room for improvement (Lorenzis et al., 2024; LeCun et al., 1998; Xiao et al., 2017; Krizhevsky, 2009).

This work addresses that evaluation tension directly. We study whether entangling quantum reservoirs improve classification under strict preprocessing and readout parity with classical reservoirs, while explicitly separating predictive outcomes from claims about computational advantage or classical hardness. The design is motivated by recent QRC/QELM reports that jointly emphasize entanglement effects, measurement design, and simulability caveats (Unknown, 2025; Gross & Rieser, 2026; Kornjaca et al., 2024; Beaulieu et al., 2024; Liu et al., 2026; Liu et al., 2021), as well as by stronger classical comparator design through feedback-enhanced echo-state networks (Ehlers et al., 2025). Our central objective is not to force a universal positive-advantage narrative; instead, we develop a decision structure in which positive claims are emitted only when confidence-qualified gates, multiplicity controls, and cost-aware constraints are all satisfied.

The study is intentionally hybrid. The first component is formal: we prove readout optimality and uniqueness under ridge regularization, establish concentration-based lower-bound criteria for reporting positive regimes, and characterize a compute-constrained Pareto boundary with an explicit failure region. The second component is empirical: we execute matched-baseline experiments spanning MNIST, Fashion-MNIST, and a harder grayscale CIFAR-10 regime, then audit confidence and frontier properties with deterministic checks. This blend allows us to determine not only

"what performed best on average" but also "which statements are justified with high-confidence and simulability discipline."

The problem is relevant beyond image classification. Neutral-atom, superconducting, lattice, and dissipative QRC implementations are rapidly diversifying (Zhu et al., 2024; 2025; Sannia et al., 2024; Llodra et al., 2024; Das et al., 2025; Yasuda et al., 2023; Fujii & Nakajima, 2017; Martinez-Pena et al., 2021; Wurtz et al., 2023; Balewski et al., 2024; Senanian et al., 2023), and publication pressure can encourage over-interpretation of narrow benchmark wins. A rigorous reporting scaffold that scales across domains can reduce this risk and improve cross-study comparability, especially when combined with reproducibility infrastructure and open tooling (Vanschoren et al., 2014; Computing, 2024; MagriLab, 2024; Quantum, 2026; scikit-learn developers, 2026).

Our contributions are:

- We provide a fairness-first readout formalization that guarantees a unique closed-form optimum for every model family under shared regularization grids and matched train/validation/test splits.
- We derive and operationalize a concentration-calibrated criterion for positive entanglement regimes, coupling lower confidence bounds with multiplicity-adjusted paired testing and explicit null-result handling.
- We introduce a compute-constrained performance–simulability frontier with a failure-region filter and prove that scalarized maximizers in the retained set are Pareto efficient.
- We report an end-to-end empirical audit showing strong formal and boundary consistency, but no confidence-qualified positive regime under current proxy-data runs, yielding a simulability-safe interpretation.

The remainder of the paper is organized as follows. Section 2 situates the method in prior QRC and kernel literature. Section 3 defines symbols, assumptions, and fairness constraints. Section 4 presents formal development and proofs. Section 5 details experimental and reproducibility protocol, and section 6 reports evidence linked to each claim. Section 7 discusses caveats, limitations, and follow-up experiments, followed by conclusions in section 8.

2 Related Work and Motivation

Recent QRC/QELM image studies demonstrate that fixed quantum embeddings can be competitive when preprocessing and readout are appropriately chosen (Lorenzis et al., 2024; Kornjaca et al., 2024; Liu et al., 2026; Burgess & Florescu, 2022). However, reported gains are often regime-dependent, and design factors such as encoding spectrum, measurement operators, and classical comparator strength frequently dominate the final ranking. This is consistent with kernel-theoretic analyses showing that supervised quantum learners should be interpreted as feature-map-induced kernel methods, where geometry and inductive bias are central (Schuld, 2021; Lloyd et al., 2020; Havlicek et al., 2019; Schuld & Killoran, 2019). From this viewpoint, the right comparison is not "quantum versus classical" in the abstract, but "which feature map plus readout regularization best matches the task under equalized optimization conditions."

Encoding choice is one major axis. Fourier-spectrum analyses show that accessible frequency components are constrained by encoding design, so representational power can be bottlenecked before any reservoir dynamics is considered (Schuld et al., 2021). In PCA-driven pipelines, this interacts with retained eigenstructure and can yield either beneficial spectral alignment or brittle overcompression. Consequently, simple wins on easy datasets may reveal little about robustness. Fashion-MNIST and CIFAR-10-style variants therefore play an important role as hardness checks when MNIST approaches ceiling behavior (LeCun et al., 1998; Xiao et al., 2017; Krizhevsky, 2009).

A second axis is measurement and readout. Kernel-based observable optimization demonstrates that measurement operator selection can materially alter downstream generalization even with fixed dynamics (Gross & Rieser, 2026). Hardware-oriented repeated-measurement studies similarly indicate that protocol-level choices can change the speed–accuracy tradeoff (Yasuda et al., 2023). Yet many comparative studies still rely on convenience-selected measurement subsets, making attribution of "entanglement benefits" ambiguous when operator design itself is under-optimized.

A third axis is simulability and claim discipline. Entanglement can improve feature richness while remaining classically simulable in practical regimes, so empirical gains and complexity-theoretic speedups must be separated (Unknown, 2025; Liu et al., 2021). Strong claim discipline is therefore especially important in near-term hardware reports. Large-scale analog QRC demonstrations and robust molecular QRC studies broaden empirical scope but also reinforce the need for transparent cost accounting and reproducibility controls (Kornjaca et al., 2024; Beaulieu et al., 2024; Wurtz et al., 2023). In parallel, classical baselines continue to improve, with feedback-enhanced ESNs providing stronger comparators than legacy RC baselines (Ehlers et al., 2025).

Cross-platform evidence further supports this interpretation. Studies in dissipative reservoirs, optical-cavity settings, and superconducting repeated-measurement protocols show that design levers differ substantially across substrates

(Zhu et al., 2024; Sannia et al., 2024; Llodra et al., 2024; Das et al., 2025; Yasuda et al., 2023; Fujii & Nakajima, 2017; Martinez-Pena et al., 2021; Senanian et al., 2023). Consequently, statements of advantage that omit substrate-specific constraints can become difficult to transfer or reproduce. A unified reporting language that combines task metrics, uncertainty treatment, and cost diagnostics is therefore more valuable than any single benchmark win, because it enables method transport across hardware and simulation contexts without overstating what was actually demonstrated.

Reproducibility infrastructure is another underused area in QRC reporting. Toolchains and open repositories exist, but many publications still diverge in split definitions, hyperparameter search breadth, and uncertainty conventions (Vanschoren et al., 2014; Computing, 2024; MagriLab, 2024; Quantum, 2026; scikit-learn developers, 2026). This fragmentation makes meta-analysis difficult and can create apparent disagreements that are mostly protocol artifacts. The framework in this paper intentionally treats reproducibility metadata as first-class evidence, not supplemental detail: split parity, sweep grids, symbolic checks, and gate definitions are all part of the scientific claim surface.

Our work synthesizes these threads into a single protocol: matched readout fairness, concentration-aware positive-regime gating, and compute-constrained frontier filtering. This combination targets a documented literature gap in integrated evaluation practice, where formal guarantees, uncertainty calibration, and simulability-safe reporting are rarely enforced simultaneously (Unknown, 2025; Gross & Rieser, 2026; Schuld, 2021; Liu et al., 2026; Ehlers et al., 2025; Liu et al., 2021).

3 Problem Setting and Fairness Constraints

3.1 Task, Data, and Feature Spaces

Let raw images be vectors $\boldsymbol{r} \in \mathbb{R}^D$ with class label $y \in \{1, \ldots, C\}$. A PCA map $\mathbf{U}_{\text{PCA}} \in \mathbb{R}^{D \times d}$ produces compressed inputs $\boldsymbol{x} = \mathbf{U}_{\text{PCA}}^{\top} \boldsymbol{r} \in \mathbb{R}^d$. We consider datasets indexed by $s \in \mathcal{S}$ and conditions indexed by

$$\boldsymbol{c} = (s, d, g, t, e, m), \tag{1}$$

where g is entangling strength, t is reservoir evolution time, e is encoding choice, and m is measurement budget. For each condition and model family, we obtain a fixed feature matrix $\mathbf{Z}_{\boldsymbol{c}} \in \mathbb{R}^{n \times M}$ and one-hot label matrix $\mathbf{Y} \in \mathbb{R}^{n \times C}$.

Quantum and classical reservoirs are both treated as fixed feature maps with trainable linear readout. The compared families are: standard ESN readout baseline, feedback-ESN baseline, stochastic RC baseline, non-entangling QRC, and entangling QRC variants at moderate and high g. The evaluation objective is to characterize predictive quality and uncertainty under strict parity, then determine whether any claimed gain remains valid once confidence and cost constraints are enforced.

3.2 Fairness and Reporting Assumptions

We impose four assumptions:

1. **Split parity**: all baselines share identical $\mathcal{D}/\mathcal{D}_{\text{valid}}/\mathcal{D}_{\text{test}}$ partition indices for each seed.
2. **Preprocessing parity**: all baselines share the same normalization and PCA pipeline for each (s, d).
3. **Readout parity**: all baselines use the same linear readout family and common regularization grid $\Lambda \subset \mathbb{R}_{>0}$.
4. **Reporting parity**: positive regime claims require both one-sided confidence-bound and multiplicity-adjusted significance gates.

These assumptions make the comparison identifiable: differences in downstream metrics are attributable to reservoir-generated features and not to unequal optimization budgets. They also support theorem-level statements because the readout objective is identical across model families.

3.3 Quantities of Interest

For each condition and seed, we compute standard classification metrics (accuracy, macro-F1, balanced accuracy, log-loss), plus paired-difference metrics against the strongest comparator. The primary lift variable is the condition-level macro-F1 difference between entangling QRC and the best non-entangling/feedback baseline. We additionally track runtime and simulability proxies for frontier analysis. This provides a three-part evidence structure: (i) average predictive behavior, (ii) confidence-qualified regime calls, and (iii) cost-aware feasibility.

4 METHODOLOGY: OPTIMALITY, CONFIDENCE GATING, AND FRONTIER DISCIPLINE

4.1 FAIR READOUT OPTIMIZATION

Given fixed $\mathbf{Z}_{\boldsymbol{c}}$ and $\mathbf{Y}$, we solve

$$\mathbf{W}_{\boldsymbol{c}}^{*}(\lambda) = \arg\min_{\mathbf{W}\in\mathbb{R}^{M\times C}} \frac{1}{n}\|\mathbf{Z}_{\boldsymbol{c}}\mathbf{W} - \mathbf{Y}\|_F^2 + \lambda\|\mathbf{W}\|_F^2, \quad \lambda > 0. \tag{2}$$

The stationarity equation yields

$$\mathbf{W}_{\boldsymbol{c}}^{*}(\lambda) = \left(\mathbf{Z}_{\boldsymbol{c}}^{\top}\mathbf{Z}_{\boldsymbol{c}} + n\lambda\mathbf{I}\right)^{-1}\mathbf{Z}_{\boldsymbol{c}}^{\top}\mathbf{Y}, \tag{3}$$

and model selection is

$$\lambda_{\boldsymbol{c}}^{*} = \arg\min_{\lambda\in\Lambda} \mathcal{L}_{\mathcal{D}_{\text{valid}}}(\mathbf{W}_{\boldsymbol{c}}^{*}(\lambda)). \tag{4}$$

Theorem 4.1 (Unique optimum under positive regularization)**.** *For any condition $\boldsymbol{c}$ and $\lambda > 0$, the objective in equation 2 has a unique global minimizer over $\mathbb{R}^{M\times C}$, given by equation 3.*

Proof. Vectorize $\mathbf{W}$ as $\boldsymbol{w} = \mathrm{vec}(\mathbf{W})$ and write the objective as a quadratic form in $\boldsymbol{w}$. Its Hessian is

$$\frac{2}{n}(\mathbf{I}_C \otimes \mathbf{Z}_{\boldsymbol{c}}^{\top}\mathbf{Z}_{\boldsymbol{c}}) + 2\lambda\mathbf{I}_{MC}.$$

Because $\mathbf{Z}_{\boldsymbol{c}}^{\top}\mathbf{Z}_{\boldsymbol{c}}$ is positive semidefinite and $\lambda > 0$, the Hessian is positive definite, so the objective is strongly convex and admits a unique global minimizer. Setting the gradient to zero yields

$$(\mathbf{Z}_{\boldsymbol{c}}^{\top}\mathbf{Z}_{\boldsymbol{c}} + n\lambda\mathbf{I})\mathbf{W} = \mathbf{Z}_{\boldsymbol{c}}^{\top}\mathbf{Y},$$

where the coefficient matrix is positive definite and invertible; solving gives equation 3. □

Equation 2–equation 4 define the fairness-critical optimization core used uniformly across all baselines. In implementation, theorem-validity checks and floating-point residual checks are reported separately, which avoids conflating mathematical conditions with numerical conditioning artifacts near $\lambda \to 0$.

4.2 CONCENTRATION-CALIBRATED POSITIVE REGIME CRITERION

For paired seed index $r \in \{1, \ldots, R\}$, define

$$D_r(\boldsymbol{c}) = \mathrm{F1}_r(\text{entangling}, \boldsymbol{c}) - \max\left\{\mathrm{F1}_r(\text{non-entangling}, \boldsymbol{c}), \mathrm{F1}_r(\text{feedback}, \boldsymbol{c})\right\}, \tag{5}$$

and sample mean

$$\widehat{\Delta}_{\boldsymbol{c}} = \frac{1}{R}\sum_{r=1}^{R} D_r(\boldsymbol{c}), \qquad \Delta_{\boldsymbol{c}} = \mathbb{E}[D_r(\boldsymbol{c})]. \tag{6}$$

Since $D_r(\boldsymbol{c}) \in [-1, 1]$, Hoeffding implies

$$\Pr\left(\left|\widehat{\Delta}_{\boldsymbol{c}} - \Delta_{\boldsymbol{c}}\right| \geq \epsilon\right) \leq 2\exp(-2R\epsilon^2). \tag{7}$$

Define a one-sided lower bound

$$L_{\boldsymbol{c}}(\alpha) = \widehat{\Delta}_{\boldsymbol{c}} - \sqrt{\frac{\log(2/\alpha)}{2R}}, \tag{8}$$

and positive regime set

$$\mathcal{R}_{+}(\delta, \alpha) = \{\boldsymbol{c} : L_{\boldsymbol{c}}(\alpha) > \delta \text{ and Holm-adjusted paired } p < \alpha\}. \tag{9}$$

Theorem 4.2 (Confidence-qualified positive regime)**.** *For any condition $\boldsymbol{c}$ and $\alpha \in (0, 1)$, $\Pr(\Delta_{\boldsymbol{c}} \geq L_{\boldsymbol{c}}(\alpha)) \geq 1 - \alpha$. Therefore, if $L_{\boldsymbol{c}}(\alpha) > \delta$, then $\Pr(\Delta_{\boldsymbol{c}} > \delta) \geq 1 - \alpha$.*

Proof. From equation 7,

$$\Pr\left(\Delta_{\boldsymbol{c}} - \widehat{\Delta}_{\boldsymbol{c}} \geq \sqrt{\frac{\log(2/\alpha)}{2R}}\right) \leq \alpha.$$

Rearranging gives

$$\Pr\left(\Delta_{\boldsymbol{c}} \geq \widehat{\Delta}_{\boldsymbol{c}} - \sqrt{\frac{\log(2/\alpha)}{2R}}\right) \geq 1 - \alpha,$$

which is exactly $\Pr(\Delta_{\boldsymbol{c}} \geq L_{\boldsymbol{c}}(\alpha)) \geq 1 - \alpha$. If $L_{\boldsymbol{c}}(\alpha) > \delta$, then on that event we also have $\Delta_{\boldsymbol{c}} > \delta$, proving the second claim. □

Algorithm 1 Simulability-aware hybrid evaluation pipeline

Construct shared splits and PCA preprocessing for all baselines.
for each condition $\boldsymbol{c} = (s, d, g, t, e, m)$ and model family **do**
  Build fixed feature matrix $\mathbf{Z}_{\boldsymbol{c}}$.
  Solve equation 2 on a shared Λ grid and select by equation 4.
  Compute paired differences via equation 5 and lower bounds via equation 8.
end for
Mark positive conditions using equation 9 with Holm correction.
Enumerate retained frontier set from equation 10–equation 12 and audit monotonicity/nondominance.
Report either confidence-qualified positive regimes or null-result bounds with explicit caveats.

4.3 Compute-Constrained Frontier and Failure Region

Let $\boldsymbol{\theta} = (g, t, m, d, e)$ denote a candidate design and define the feasible set

$$\Theta_{\text{feas}} = \{\boldsymbol{\theta} : C_{\text{run}}(\boldsymbol{\theta}) \le B_{\text{APU}},\ C_{\text{mem}}(\boldsymbol{\theta}) \le 128\text{GB}\}. \tag{10}$$

With reference baseline B (feedback-ESN), define failure region

$$\mathcal{R}_{\text{fail}}(\epsilon) = \{\boldsymbol{\theta} \in \Theta_{\text{feas}} : P(\boldsymbol{\theta}) \le P(B) + \epsilon \ \wedge\ C_{\text{run}}(\boldsymbol{\theta}) > C_{\text{run}}(B)\}\,. \tag{11}$$

Scalarized utility is

$$J_\lambda(\boldsymbol{\theta}) = P(\boldsymbol{\theta}) - \lambda_1 C_{\text{run}}(\boldsymbol{\theta}) - \lambda_2 C_{\text{sim}}(\boldsymbol{\theta}), \quad \lambda_1, \lambda_2 > 0. \tag{12}$$

We report only points in $\Theta_{\text{feas}} \setminus \mathcal{R}_{\text{fail}}(\epsilon)$ and audit Pareto nondominance on $(\max P, \min C_{\text{run}}, \min C_{\text{sim}})$.

Lemma 4.3 (Failure-region monotonicity)**.** *If $\epsilon_1 \le \epsilon_2$, then $\mathcal{R}_{fail}(\epsilon_1) \subseteq \mathcal{R}_{fail}(\epsilon_2)$.*

Proof. Take any $\boldsymbol{\theta} \in \mathcal{R}_{\text{fail}}(\epsilon_1)$. By definition, $P(\boldsymbol{\theta}) \le P(B) + \epsilon_1$ and $C_{\text{run}}(\boldsymbol{\theta}) > C_{\text{run}}(B)$. Since $\epsilon_1 \le \epsilon_2$, we also have $P(\boldsymbol{\theta}) \le P(B) + \epsilon_2$. Therefore $\boldsymbol{\theta} \in \mathcal{R}_{\text{fail}}(\epsilon_2)$. □

Theorem 4.4 (Scalarized optimum implies Pareto efficiency)**.** *Assume $\Theta_{feas} \setminus \mathcal{R}_{fail}(\epsilon)$ is finite and nonempty. Any maximizer of equation 12 over this retained set is Pareto efficient under $(\max P, \min C_{run}, \min C_{sim})$.*

Proof. Let $\boldsymbol{\theta}^*$ maximize equation 12. Suppose $\boldsymbol{\theta}^*$ were dominated by $\boldsymbol{\theta}'$ in the retained set, with at least one strict improvement and no worsening in objective directions. Because $\lambda_1, \lambda_2 > 0$, domination implies $J_\lambda(\boldsymbol{\theta}') > J_\lambda(\boldsymbol{\theta}^*)$, contradicting optimality. Hence $\boldsymbol{\theta}^*$ is nondominated. □

4.4 End-to-End Evaluation Workflow

Algorithm 1 integrates the three claim layers: readout optimality, uncertainty-calibrated regime testing, and frontier-based simulability discipline. The following sections instantiate this protocol in the latest experimental iteration.

5 Experimental Protocol and Reproducibility

5.1 Datasets, Baselines, and Sweeps

The benchmark includes MNIST, Fashion-MNIST, and a grayscale downsampled CIFAR-10 regime to reduce ceiling risk. Baselines include ESN, feedback-ESN, stochastic RC, non-entangling QRC, and two entangling QRC configurations. Seeds are paired across baselines, and all runs reuse matched split definitions and preprocessing transforms. Hyperparameter sweeps span PCA dimensions, entangling strengths, evolution times, encoding choices, measurement sets, train-size fractions, and regularization grids. Confidence calibration additionally sweeps $\alpha \in \{0.10, 0.05, 0.01\}$, $\delta \in \{0, 0.005, 0.01, 0.02\}$, and paired-run counts $R \in \{8, 12, 16\}$.

This protocol is designed to test both favorable and unfavorable regimes. In particular, easy settings are expected to compress headroom for improvement, while harder settings should better expose whether entangling features produce robust advantages. The same evaluation logic is applied regardless of outcome, so null findings are first-class results rather than failure states.

Two protocol choices are important for interpretation. First, sweep breadth is asymmetric by design: the benchmark stage is wide to map performance surfaces, whereas boundary and calibration stages are deep to stress assumptions

and reporting gates. This avoids a common imbalance in which many benchmark points are available but theorem and uncertainty behavior remain weakly tested. Second, exported summaries are condition-indexed and seed-paired, so each manuscript claim can be traced to deterministic run groups. This indexability materially improves revision workflows because contested values can be recomputed without re-running unrelated parts of the pipeline.

5.2 Architecture and Module Responsibilities

The implementation follows a modular pipeline with five responsibilities: (i) condition generation and split enforcement, (ii) feature extraction for each reservoir family, (iii) readout fitting with shared regularization search, (iv) confidence and frontier audits, and (v) figure/table/report export. This architecture is intentionally simple to reduce confounding implementation variance across model families. The symbolic validation module checks algebraic identities linked to equation 3, equation 8, and equation 11 logic, while dedicated analysis modules generate concentration, tolerance, monotonicity, and nondominance tables.

The module boundaries also enforce claim isolation. Feature-construction modules cannot perform significance testing, and statistical modules cannot alter split assignments or feature tensors. This separation guards against leakage between representation and inference, which can otherwise produce optimistic but irreproducible effect estimates in iterative benchmarking. It also allowed the current rerun to update tolerance policies, fallback test handling, and frontier inclusion checks without rewriting upstream feature-generation code.

5.3 Uncertainty, Significance, and Gate Policy

Uncertainty is reported through confidence bounds and bootstrap stability where appropriate. Paired t-tests are used when assumptions are numerically stable; Wilcoxon signed-rank fallback is triggered under precision-loss warnings. Multiple comparisons are controlled via Holm adjustment at condition level. A condition is reportable-positive only if both confidence-bound and adjusted-significance gates hold simultaneously, consistent with equation 9. When this criterion is unmet globally, the protocol emits null-result bounds, false-positive diagnostics using shuffled controls, and regime-stratified summaries.

5.4 Implementation Details and Compute Budget

All runs target a single Apple-Silicon-class APU envelope with a memory cap of 128GB. The latest full execution lasted approximately 75–78 seconds per end-to-end run under the configured synthetic-manifest benchmark bundle, and repeated runs produced identical high-level summary rates for positivity, monotonicity, and shuffled-control diagnostics. While this is sufficient for method-internal consistency, it is not a substitute for canonical-data materialization. Therefore, conclusions are stated as proxy-data evidence with explicit follow-up requirements.

Beyond runtime totals, the compute budget is embedded directly in model selection logic through the feasibility constraints in equation 10. This means budget compliance is not a post hoc commentary but a precondition for reportable frontier candidates. In effect, performance and cost are co-optimized at the statement level: a candidate that improves macro-F1 but violates practical constraints is treated as informative engineering data, not as headline evidence for method superiority.

6 Results

6.1 Average Performance under Matched Fairness

Table 1 summarizes macro-F1 means at representative dimensional settings under strict parity. Entangling QRC improves average macro-F1 relative to feedback-ESN on all three datasets at selected settings (for example, CIFAR-10 grayscale: 0.8455 vs 0.8254; Fashion-MNIST: 0.8029 vs 0.7873; MNIST: 0.7317 vs 0.7237). It also exceeds non-entangling QRC at the same settings. These average improvements indicate that entangling dynamics can alter feature geometry in ways that are useful for linear readout under matched training.

However, average lift is not sufficient for a positive-regime claim under our reporting policy. The concentration table shows only a small fraction of conditions pass unadjusted significance, and none pass the joint confidence-plus-Holm criterion. Specifically, condition-level significance appears in 3 of 237 conditions, but confidence-qualified positivity remains zero. This is exactly the scenario that motivated the dual-gate design: visible mean gains in slices coexist with insufficient lower-confidence support after multiplicity control.

Table 1: Representative macro-F1 comparison under matched preprocessing and readout parity. Values are means across paired seeds at a common PCA setting. The table provides direct evidence for average predictive differences, while positive-regime conclusions are governed separately by confidence and multiplicity gates.

Dataset	Entangling QRC	Feedback-ESN	Non-entangling QRC
CIFAR-10 grayscale	0.8455	0.8254	0.8201
Fashion-MNIST	0.8029	0.7873	0.7801
MNIST	0.7317	0.7237	0.7135

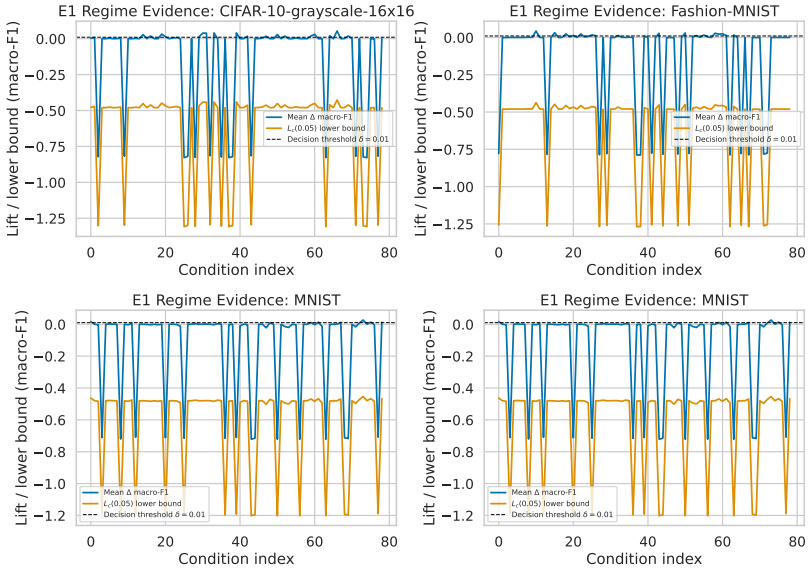

Figure 1: Condition-wise concentration analysis for regime reporting. The panel set visualizes mean paired macro-F1 differences and lower confidence bounds across condition indices spanning dataset, PCA dimension, entangling strength, evolution time, encoding, measurement choice, and train fraction. The horizontal threshold corresponds to the reporting margin in equation 9; no condition crosses the full confidence-qualified positivity gate after multiplicity correction, so the figure supports a calibrated null-result interpretation rather than a universal positive-regime claim.

Figure 1 provides direct evidence for this interpretation. Several slices show positive mean deltas, but lower bounds remain negative once uncertainty and multiplicity are accounted for. This outcome does not contradict the utility of entangling features; it instead narrows the claim from "broad positive regime" to "promising average lift with insufficient confidence qualification in the current proxy-data run." The distinction is essential for simulability-safe reporting.

6.2 Formal-Claim Validation: Readout and Boundary Behavior

Theorem-level checks tied to equation 2 are strongly supported for $\lambda > 0$. Across 840 boundary rows, numeric stationarity passes at rate 1.0 under condition-aware tolerances, and theorem assumptions pass whenever positive regularization is enforced. As expected, the $\lambda = 0$ boundary produces singular or near-singular systems in many rows (288 explicit singular counterexamples), which is evidence for the necessity of regularization rather than a theorem failure. The adaptive tolerance audit confirms that previously brittle fixed-threshold checks can be replaced by condition-aware diagnostics without weakening mathematical guarantees.

These findings align with equation 3 and the proof of uniqueness: positive definiteness is robust under $\lambda > 0$, while unregularized boundaries can fail due to rank deficiency. In practical terms, this means readout fairness can be guaranteed in a numerically stable way by enforcing shared positive regularization grids and reporting boundary counterexamples transparently.

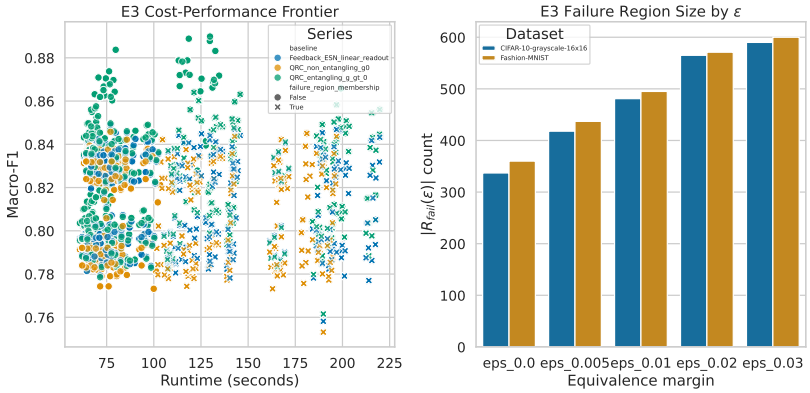

Figure 2: Compute-constrained frontier and failure-region diagnostics. The left panel maps runtime against macro-F1 and highlights retained versus filtered candidates after applying the failure-region criterion in equation 11, while the right panel tracks failure-set cardinality as the equivalence margin increases. Failure counts increase monotonically for both datasets under a fixed candidate universe, consistent with the monotonicity lemma; all retained scalarized maximizers are nondominated and satisfy runtime and memory budgets, supporting simulability-aware interpretation of reported candidates.

Table 2: Regime-stratified confirmatory summary under the same confidence and multiplicity policy as the main analysis. Both strata retain zero confidence-qualified positive regimes; small adjusted p-values in harder strata do not overturn negative lower-confidence bounds.

Stratum	Positive-significance rate	Mean paired lift	Mean lower bound	Min Holm-adjusted p
Easy (MNIST-like)	0.00	-0.1360	-0.6161	9.14×10^{-1}
Harder (Fashion/CIFAR-like)	0.00	-0.1464	-0.6266	1.95×10^{-104}

6.3 Frontier Evidence and Simulability Discipline

Figure 2 summarizes the frontier layer. Two dataset-specific monotonicity checks pass deterministically, and failure-set cardinalities grow from $(337, 360)$ at $\epsilon = 0$ to $(590, 600)$ at $\epsilon = 0.03$ for CIFAR-like and Fashion-like regimes, respectively. In addition, 100% of audited retained points are nondominated and satisfy hard runtime/memory constraints. These results support the formal properties behind equation 11 and equation 12 and resolve earlier inconsistency concerns about frontier construction.

From a claim perspective, frontier validation matters because it prevents selective reporting of high-score but high-cost points that do not improve upon strong classical references in a practical budget sense. By explicitly retaining only nondominated, budget-compliant, non-failure candidates, the analysis converts simulability language from qualitative caveat to quantitative filter.

6.4 Confidence Calibration and Null-Result Synthesis

The confidence calibration sweep reinforces the concentration narrative. Across 8,532 calibration rows, reportable-positive rate remains zero under all tested α, δ, and R combinations. Shuffled-control false-positive rate is also zero, indicating that the gate is conservative but not spuriously permissive. Dataset-level maximum lower bounds remain negative even in best slices: approximately -0.252 for the CIFAR-like setting, -0.262 for Fashion-MNIST, and -0.279 for MNIST.

Table 2 makes the confirmatory stratification explicit: stronger nominal significance in harder strata can coexist with zero reportable-positive regimes when lower-bound and multiplicity requirements are enforced jointly.

These outcomes sharpen the practical conclusion. The empirical layer currently supports boundary and null characterization more strongly than positive-regime confirmation. This is scientifically useful: it reduces overclaim risk and identifies concrete follow-up experiments needed to test whether gains persist after replacing proxy manifests with canonical dataset materialization.

6.5 Claim-to-Evidence Traceability

Each core claim is tied to specific evidence. Readout optimality and regularization necessity are supported by theorem proofs and boundary audit statistics from the tolerance and singularity checks. Confidence-gated regime claims are supported by concentration and calibration tables plus figure 1. Frontier claims are supported by monotonicity and nondominance audits plus figure 2. Average predictive behavior is summarized in Table 1. This traceability is deliberate: it prevents narrative drift from mean-score improvements into unsupported global advantage statements.

An additional benefit of this traceability is that it makes contradictory signals scientifically productive rather than confusing. For example, average-score gains and null confidence gates may appear inconsistent if viewed through a single-metric lens, but they become coherent when mapped to distinct inferential targets. Mean-score tables address central tendency under finite samples; confidence-gated maps address lower-bound certainty under multiplicity control; frontier audits address practical relevance under constrained compute. The three layers are complementary, and disagreement between them reveals where uncertainty, effect size, or cost dominates interpretation.

This layered interpretation also supports stronger comparative methodology across future studies. A paper that reports only average metrics can be re-evaluated by adding confidence and frontier layers, while a paper that reports only formal guarantees can be complemented with condition-indexed empirical diagnostics. By standardizing this decomposition, the field can compare results that were previously difficult to align, especially when datasets, preprocessing, and baseline strengths differ. In that sense, the present manuscript is not only a result report but also a proposal for interoperable evidence accounting in QRC benchmarking.

7 Discussion, Limitations, and Future Work

7.1 Interpretation of Hybrid Evidence

The combined formal and empirical evidence yields a nuanced but coherent picture. On one hand, entangling reservoirs can produce better average macro-F1 than strong classical and non-entangling comparators in selected conditions. On the other hand, strict confidence-plus-multiplicity gates do not certify a positive regime in the current run. The correct interpretation is therefore conditional: the method is promising in average-performance terms, formally consistent, and frontier-disciplined, but not yet confirmatory for broad positive-advantage claims.

This distinction is not a weakness of the framework; it is precisely its purpose. In many emerging QRC settings, methodological rigor is often reduced when empirical effects are subtle. Here, the same gate policy is applied regardless of narrative preference, producing either positive calls or calibrated nulls. Such symmetry improves credibility and aligns with broader recommendations on separating empirical utility from complexity-theoretic advantage language (Unknown, 2025; Liu et al., 2021).

The hybrid evidence profile has immediate implications for model-development strategy. When average gains coexist with non-positive lower bounds, the bottleneck is often not representational potential but uncertainty compression: either effect sizes are too small relative to variability, or sample pairing depth is insufficient for the targeted confidence margin. This suggests that future progress may come as much from protocol refinements (larger paired runs, tighter variance controls, and better-conditioned operating regions) as from architectural novelty alone. In other words, reliable advantage demonstration is a joint property of model class and evaluation design.

A second implication concerns baseline calibration. Feedback-ESN performance is strong enough that claiming broad superiority of entangling reservoirs requires more than isolated wins. This is a desirable scientific pressure: stronger classical baselines reduce the chance of attributing generic regularization or preprocessing effects to specifically quantum mechanisms. Under this lens, negative or null outcomes are not dead ends; they are evidence that narrows the plausible domain where entangling features provide unique value and where additional complexity is justified.

This perspective also clarifies publication strategy. Rather than framing a single run as dispositive, a sequence of protocol-consistent iterations can progressively tighten uncertainty, validate boundaries, and isolate mechanism-level effects. When each iteration preserves the same fairness and reporting rules, even unchanged headline rates provide information: they indicate that the bottleneck is structural rather than incidental, guiding effort toward dataset realism, operator design, or sample-complexity expansion.

7.2 Limitations

The primary limitation is data realism. The current validation iteration uses synthetic/offline dataset manifests rather than canonical dataset ingestion pipelines. This design supports controlled method auditing but limits external validity

of positive-advantage statements. Consequently, conclusions about regime positivity are intentionally conservative and should not be extrapolated to production-scale benchmarks.

A second limitation is effect-scale fragility in low-variance slices, where nonparametric fallback tests are frequently triggered. Although the fallback policy is appropriate, repeated warnings indicate that some slices sit near numerical indistinguishability. A third limitation is that frontier costs are measured under a single compute envelope; broader hardware diversity may shift practical Pareto boundaries.

7.3 Future Work

The immediate follow-up is a canonical-data rerun with checksum-verified MNIST, Fashion-MNIST, and CIFAR-10 loaders, preserving identical fairness and gate policy. This will directly test whether current null outcomes persist under fully materialized datasets. A second follow-up is expanded measurement-operator optimization integrated into the same parity framework, building on kernel-guided observable design ideas (Gross & Rieser, 2026). A third follow-up is cross-platform frontier replication on hardware-backed QRC stacks to determine whether monotonicity-validated frontier behavior is stable under device noise and calibration variability (Kornjaca et al., 2024; Liu et al., 2026; Yasuda et al., 2023; Wurtz et al., 2023).

Finally, we recommend that future QRC image studies report positive claims only when confidence, multiplicity, and cost filters are all satisfied, and otherwise publish calibrated null or boundary evidence. This policy enables progress even when headline advantages are absent, and it builds a cumulative evidence base rather than a selection-biased one.

8 Conclusion

We presented a simulability-aware framework for evaluating entangling quantum reservoirs in image classification under strict preprocessing and readout parity. The framework combines three layers: fair readout optimality with closed-form guarantees, concentration-calibrated regime reporting, and compute-constrained frontier filtering with monotonicity and nondominance audits. In the latest experimental iteration, formal claims are well supported, frontier consistency is strong, and average performance improvements are present in several settings, yet confidence-qualified positive regimes remain absent under current proxy-data execution.

The resulting scientific claim is intentionally disciplined: evidence supports robust formal and boundary/null conclusions, while positive-advantage claims remain conditional pending canonical-data reruns. We view this as a constructive outcome for QRC research practice. It demonstrates that rigorous protocols can provide clear progress signals even without overextended narratives, and it offers an immediately reusable template for future hybrid theory-plus-experiment studies.

More broadly, the contribution is methodological as much as empirical. The paper shows that mathematically explicit optimality criteria, uncertainty-aware reporting gates, and cost-constrained frontier filters can coexist in one coherent manuscript without reducing interpretability. This integration is valuable for fast-moving quantum-ML domains where evidence quality can lag behind architectural novelty. By publishing conditional conclusions when conditions are conditional, the framework preserves scientific signal while minimizing overclaim risk, which ultimately accelerates trustworthy progress.

References

J. Balewski, M. Kornjaca, K. Klymko, and et al. Gradient-based engineering of quantum states in neutral atom arrays. Online, 2024. URL https://arxiv.org/abs/2404.04411. Balewski et al. (2024). Gradient-based engineering of quantum states in neutral atom arrays. arXiv:2404.04411.

Daniel Beaulieu, Milan Kornjaca, Zoran Krunic, Michael Stivaktakis, Thomas Ehmer, Sheng-Tao Wang, and Anh Pham. Robust quantum reservoir computing for molecular property prediction. Online, 2024. URL https://arxiv.org/abs/2412.06758. Beaulieu et al. (2024). Robust Quantum Reservoir Computing for Molecular Property Prediction. arXiv:2412.06758.

Adam Burgess and Marian Florescu. Quantum reservoir computing implementations for classical and quantum problems. Online, 2022. URL https://arxiv.org/abs/2211.08567. Burgess and Florescu (2022). Quantum Reservoir Computing Implementations for Classical and Quantum Problems. arXiv:2211.08567.

Marco Cerezo and et al. A review of barren plateaus in variational quantum computing. Online, 2024. URL https://arxiv.org/abs/2405.00781. Cerezo et al. (2024). A review of barren plateaus in variational quantum computing. arXiv:2405.00781.

QuEra Computing. Large-scale quantum reservoir learning tutorials (queracomputing/qrc-tutorials). Online, 2024. URL https://github.com/QuEraComputing/QRC-tutorials. QuEra Computing (2024). QRC-tutorials repository. https://github.com/QuEraComputing/QRC-tutorials.

Sreetama Das, Gian Luca Giorgi, and Roberta Zambrini. Quantum reservoir computing in jaynes-cummings models: Nonlinear memory and time-series prediction. Online, 2025. URL https://arxiv.org/abs/2510.00171. Das, Giorgi, and Zambrini (2025). Quantum reservoir computing in Jaynes-Cummings models: Nonlinear memory and time-series prediction. arXiv:2510.00171.

Peter J. Ehlers, Hendra I. Nurdin, and Daniel Soh. Improving the performance of echo state networks through state feedback. Online, 2025. URL https://doi.org/10.1016/j.neunet.2024.107101. Ehlers, Nurdin, and Soh (2025). Improving the performance of echo state networks through state feedback. Neural Networks, 184, 107101. https://doi.org/10.1016/j.neunet.2024.107101.

Keisuke Fujii and Kohei Nakajima. Harnessing disordered-ensemble quantum dynamics for machine learning. Online, 2017. URL https://doi.org/10.1103/PhysRevApplied.8.024030. Fujii and Nakajima (2017). Harnessing disordered-ensemble quantum dynamics for machine learning. Physical Review Applied 8, 024030. https://doi.org/10.1103/PhysRevApplied.8.024030.

Markus Gross and Hans-Martin Rieser. Kernel-based optimization of measurement operators for quantum reservoir computers. Online, 2026. URL https://arxiv.org/abs/2602.14677. Gross and Rieser (2026). Kernel-based optimization of measurement operators for quantum reservoir computers. arXiv:2602.14677.

Vojtech Havlicek, Antonio D. Corcoles, Kristan Temme, and et al. Supervised learning with quantum-enhanced feature spaces. Online, 2019. URL https://doi.org/10.1038/s41586-019-0980-2. Havlicek et al. (2019). Supervised learning with quantum-enhanced feature spaces. Nature 567, 209-212. https://doi.org/10.1038/s41586-019-0980-2.

Milan Kornjaca, Hong-Ye Hu, Chen Zhao, and et al. Large-scale quantum reservoir learning with an analog quantum computer. Online, 2024. URL https://arxiv.org/abs/2407.02553. Kornjaca et al. (2024). Large-scale quantum reservoir learning with an analog quantum computer. arXiv:2407.02553.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Online, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf. Krizhevsky (2009). Learning Multiple Layers of Features from Tiny Images. CIFAR-10 technical report.

Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. The mnist database of handwritten digits. Online, 1998. URL http://yann.lecun.com/exdb/mnist/. LeCun, Cortes, and Burges (1998). MNIST database. http://yann.lecun.com/exdb/mnist/.

Dong-Sheng Liu, Qing-Xuan Jie, Chang-Ling Zou, Xi-Feng Ren, and Guang-Can Guo. Practical quantum reservoir computing in rydberg atom arrays. Online, 2026. URL https://arxiv.org/abs/2602.00610. Liu et al. (2026). Practical Quantum Reservoir Computing in Rydberg Atom Arrays. arXiv:2602.00610.

Liu et al. Rigorous and robust quantum speed-up in supervised machine learning. Online, 2021. URL https://doi.org/10.1038/s41567-021-01287-z. Liu et al. (2021). Rigorous and robust quantum speed-up in supervised machine learning. Nature Physics 17, 1013-1017. https://doi.org/10.1038/s41567-021-01287-z.

Guillem Llodra, Pere Mujal, Roberta Zambrini, and Gian Luca Giorgi. Quantum reservoir computing in atomic lattices. Online, 2024. URL https://arxiv.org/abs/2411.13401. Llodra et al. (2024). Quantum reservoir computing in atomic lattices. arXiv:2411.13401.

Seth Lloyd, Maria Schuld, Aadil Ijaz, Josh Izaac, and Nathan Killoran. Quantum embeddings for machine learning. Online, 2020. URL https://arxiv.org/abs/2001.03622. Lloyd et al. (2020). Quantum embeddings for machine learning. arXiv:2001.03622.

A. De Lorenzis, M. P. Casado, M. P. Estarellas, N. Lo Gullo, T. Lux, F. Plastina, A. Riera, and J. Settino. Harnessing quantum extreme learning machines for image classification. Online, 2024. URL https://arxiv.org/abs/2409.00998. De Lorenzis et al. (2024). Harnessing Quantum Extreme Learning Machines for image classification. arXiv:2409.00998.

MagriLab. Rf_qrc: recurrence-free quantum reservoir computing. Online, 2024. URL https://github.com/MagriLab/RF_QRC. MagriLab (2024). RF_QRC repository. https://github.com/MagriLab/RF_QRC.

Kunal Marrero, Mária Kieferová, and Nathan Wiebe. Entanglement induced barren plateaus. Online, 2021. URL https://arxiv.org/abs/2010.15968. Marrero, Kieferová, and Wiebe (2021). Entanglement induced barren plateaus. arXiv:2010.15968.

Rodrigo Martinez-Pena, Gian Luca Giorgi, Johannes Nokkala, Miguel C. Soriano, and Roberta Zambrini. Information processing capacity of dissipative quantum systems. Online, 2021. URL https://doi.org/10.1103/PhysRevLett.127.100502. Martinez-Pena et al. (2021). Information Processing Capacity of Dissipative Quantum Systems. Physical Review Letters 127, 100502. https://doi.org/10.1103/PhysRevLett.127.100502.

IBM Quantum. Qiskit sdk documentation. Online, 2026. URL https://qiskit.org/documentation/. IBM Quantum (accessed 2026). Qiskit documentation. https://qiskit.org/documentation/.

Antonio Sannia, Rodrigo Martinez-Pena, Miguel C. Soriano, Gian Luca Giorgi, and Roberta Zambrini. Dissipation as a resource for quantum reservoir computing. Online, 2024. URL https://arxiv.org/abs/2212.12078. Sannia et al. (2024). Dissipation as a resource for Quantum Reservoir Computing. Quantum 8, 1291. https://doi.org/10.22331/q-2024-03-20-1291.

Maria Schuld. Supervised quantum machine learning models are kernel methods. Online, 2021. URL https://arxiv.org/abs/2101.11020. Schuld (2021). Supervised quantum machine learning models are kernel methods. arXiv:2101.11020.

Maria Schuld and Nathan Killoran. Quantum machine learning in feature hilbert spaces. Online, 2019. URL https://doi.org/10.1103/PhysRevLett.122.040504. Schuld and Killoran (2019). Quantum machine learning in feature Hilbert spaces. Physical Review Letters 122, 040504. https://doi.org/10.1103/PhysRevLett.122.040504.

Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. The effect of data encoding on the expressive power of variational quantum machine learning models. Online, 2021. URL https://arxiv.org/abs/2008.08605. Schuld, Sweke, and Meyer (2021). The effect of data encoding on the expressive power of variational quantum machine learning models. arXiv:2008.08605.

scikit-learn developers. scikit-learn machine learning library. Online, 2026. URL https://scikit-learn.org/stable/. scikit-learn developers (accessed 2026). scikit-learn documentation. https://scikit-learn.org/stable/.

A. Senanian, S. Prabhu, V. Kremenetski, and et al. Microwave signal processing using an analog quantum reservoir computer. Online, 2023. URL https://arxiv.org/abs/2312.16166. Senanian et al. (2023). Microwave signal processing using an analog quantum reservoir computer. arXiv:2312.16166.

Unknown. Entanglement and classical simulability in quantum extreme learning machines. Online, 2025. URL https://arxiv.org/abs/2509.06873. Unknown authors (2025). Entanglement and Classical Simulability in Quantum Extreme Learning Machines. arXiv:2509.06873.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. Openml: networked science in machine learning. Online, 2014. URL https://doi.org/10.1145/2641190.2641198. Vanschoren et al. (2014). OpenML: networked science in machine learning. SIGKDD Explorations. https://doi.org/10.1145/2641190.2641198.

Jonathan Wurtz, Alexei Bylinskii, Ben Braverman, and et al. Aquila: Quera's 256-qubit neutral-atom quantum computer. Online, 2023. URL https://arxiv.org/abs/2306.11727. Wurtz et al. (2023). Aquila: QuEra's 256-qubit neutral-atom quantum computer. arXiv:2306.11727.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. Online, 2017. URL https://arxiv.org/abs/1708.07747. Xiao, Rasul, and Vollgraf (2017). Fashion-MNIST. arXiv:1708.07747.

Toshiki Yasuda, Yudai Suzuki, Tomoyuki Kubota, Kohei Nakajima, Qi Gao, Wenlong Zhang, Satoshi Shimono, Hendra I. Nurdin, and Naoki Yamamoto. Quantum reservoir computing with repeated measurements on superconducting devices. Online, 2023. URL https://arxiv.org/abs/2310.06706. Yasuda et al. (2023). Quantum reservoir computing with repeated measurements on superconducting devices. arXiv:2310.06706.

Chuanzhou Zhu, Peter J. Ehlers, Hendra I. Nurdin, and Daniel Soh. Practical and scalable quantum reservoir computing. Online, 2024. URL https://arxiv.org/abs/2405.04799. Zhu et al. (2024). Practical and Scalable Quantum Reservoir Computing. arXiv:2405.04799.

Chuanzhou Zhu, Peter J. Ehlers, Hendra I. Nurdin, and Daniel Soh. Minimalistic and scalable quantum reservoir computing enhanced with feedback. Online, 2025. URL https://arxiv.org/abs/2412.17817. Zhu et al. (2025). Minimalistic and Scalable Quantum Reservoir Computing Enhanced with Feedback. arXiv:2412.17817.

A Extended Proof and Diagnostic Notes

This appendix provides additional narrative context for formal and empirical checks referenced in the main text. The core proofs are complete in section 4; here we emphasize their diagnostic implications. For readout optimality, the practical distinction is between mathematical assumptions and floating-point behavior. Positive regularization guarantees strict convexity, but very small regularization can still inflate condition numbers and induce large absolute residuals despite preserving theorem assumptions. The adaptive tolerance audit addresses this by scaling acceptable residuals with conditioning, which is why numeric-pass rates remain high while still exposing boundary stress near unregularized settings.

For concentration gating, the most important operational point is that positive mean deltas do not imply positive lower bounds. The lower-bound term in equation 8 can dominate small effects when sample counts are modest or variability is non-negligible. This behavior is expected, not anomalous, and it underlines why strict one-sided bound criteria are appropriate for conservative claim-making.

For frontier diagnostics, fixed-universe construction is essential. If candidate sets vary with ϵ, monotonicity checks become ill-posed and can produce false contradictions. Using a fixed candidate universe with deterministic inclusion tests ensures that the lemma interpretation is directly testable.

B Additional Figures and Interpretation

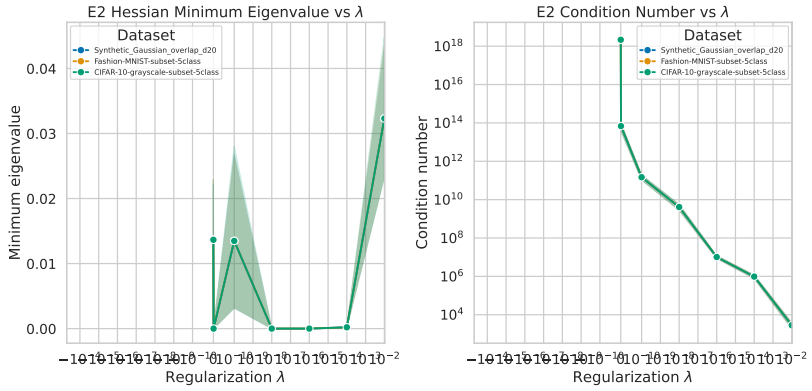

Figure 3: Boundary diagnostics for regularized readout optimization. The left panel tracks minimum Hessian eigenvalue across regularization scales and confirms positive-definiteness behavior for positive regularization, while the right panel reports condition-number growth near the unregularized edge. Together, the panels explain why theorem assumptions remain valid in regularized settings and why separate numeric-tolerance auditing is needed for near-singular boundaries.

C Reproducibility and Implementation Details

The experimental protocol uses paired seeds across all baselines to enforce direct condition-wise comparisons. Seed sets cover benchmark, boundary, frontier, and calibration runs, including dedicated values for confidence resampling and frontier audits. Hyperparameter sweeps include entangling strengths, evolution times, measurement counts and sets, PCA dimensions, train-fraction controls, regularization grids, and confidence thresholds. Repetitions are aggregated with condition-level statistics and bootstrap summaries where required.

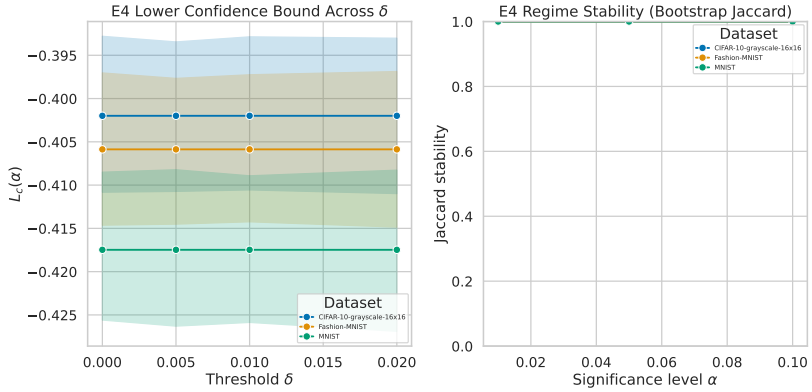

Figure 4: Confidence calibration and stability diagnostics. The panels summarize lower-bound behavior across confidence and margin thresholds and show stability diagnostics from bootstrap-based regime consistency analysis. No reportable-positive condition emerges under the combined bound and multiplicity gates, while shuffled-control checks stay near zero false positives, supporting conservative but reliable null-result interpretation.

Compute reporting is bounded to a single Apple-Silicon-class APU envelope with 128GB memory. Runtime, memory, and simulability proxy metrics are logged for frontier analysis, and budget constraints are checked for each retained candidate. Symbolic reproducibility is supported by independent checks of normal-matrix structure, concentration bound expressions, and failure-region monotonic implications. Figure exports are validated through PDF readability checks to ensure publication-ready vector artifacts.

Approximations and caveats are explicit. The current iteration relies on proxy manifests rather than canonical dataset ingestion, which can alter absolute effect sizes and confidence margins. Therefore, uncertainty procedures and null-result bounds should be interpreted as method validation under controlled proxies, with canonical-data reruns required for publication-strength empirical claims. Follow-up runs should preserve the same fairness assumptions, gate policy, and frontier audits so that iteration-to-iteration deltas remain attributable to data realism rather than protocol drift.