

QuantumNovelty Review Showcase

Five real quantum-computing papers PRX Quantum, npj Quantum Information, Quantum,
and one arXiv preprint

Generated by the QuantumNovelty paper-audit pipeline

June 2026

Project	QuantumNovelty (QN) — audit-and-falsify framework for quantum-computing research	
Repository	https://github.com/boltzmannentropy/QuantumNovelty	
Author	Shlomo Kashani (QNeura.ai)	
LLM backend	Claude Code CLI (2.1.1 (Claude Code))	
Model snapshots used	claude-haiku-4-5-20251001	claude-opus-4-5-20251101
Report generated	2026-06-10 23:18 by build_reviews.py	

Disclaimer. These reviews are generated end-to-end by AI (the QuantumNovelty agent pipeline running on the Claude Code CLI). They are provided strictly for academic and demonstration purposes — to show what the framework produces on real papers. We make NO claims about the correctness, quality, novelty, or publication-worthiness of the papers under review; four passed real peer review at their respective journals and one is a public arXiv preprint, and nothing here should be read as criticism of the authors or as a substitute for human peer review. Paper copyrights remain with their authors and publishers.

1 Papers under review

Tag	Paper	Venue	arXiv
1cutrotter	Simple and high-precision Hamiltonian simulation by compensating Trotter error with linear combination of unitary operations Pei Zeng, Jinzhao Sun, Liang Jiang, Qi Zhao	PRX Quantum 6, 010359 (2025)	2212.04566
flowvqe	Generative flow-based warm start of the variational quantum eigensolver Hang Zou, Martin Rahm, Anton Frisk Kockum, Simon Olsson	npj Quantum Information (2025)	2507.01726
hwqml	Trainability and Expressivity of Hamming-Weight Preserving Quantum Circuits for Machine Learning Léo Monbroussou, Elliott Z. Mamon, Jonas Landman, Alex B. Grilo, Romain Kukla, Elham Kashefi	Quantum 9, 1745 (2025)	2309.15547
qcnn	Quantum Convolutional Neural Networks are Effectively Classically Simulable Pablo Bermejo, Paolo Braccia, Manuel S. Rudolph, Zoë Holmes, Lukasz Cincio, M. Cerezo	PRX Quantum 7, 020304 (2026)	2408.12739
majorana2	20 Second Parity Lifetime in an InAs-Pb Tetron Device Microsoft Quantum	arXiv:2606.03884 (preprint, June 2026)	2606.03884

The peer-reviewed papers cover every target venue in QN’s journal registry; `hwqml` and `qcnn` cover quantum machine learning, and `majorana2` is the hardware entry — a preprint audited against the PRX Quantum rubric.

2 Token + cost ledger

Every LLM call records the model snapshot ID, exact input/output token counts, and USD cost from the Claude CLI’s JSON envelope.

Paper	Stage	Input tk	Output tk	Cost (\$)	Elapsed (s)
lcutrotter	Deep-research review	1	1,336	\$0.3790	36.5
lcutrotter	Reviewer panel 5 voices	1	5,293	\$0.7757	122.0
lcutrotter	Logical-fallacy report	1	2,348	\$0.7012	45.4
lcutrotter	Stage-6 CQE narrative	2	1,195	\$0.0806	31.9
lcutrotter total		5	10,172	\$1.9365	235.8
flowvqe	Deep-research review	1	1,318	\$0.2581	38.0
flowvqe	Reviewer panel 5 voices	1	4,646	\$0.3421	109.5
flowvqe	Logical-fallacy report	1	2,863	\$0.2955	52.8
flowvqe	Stage-6 CQE narrative	2	1,180	\$0.0721	31.7
flowvqe total		5	10,007	\$0.9678	232.0
hwqml	Deep-research review	1	1,546	\$0.4351	42.7
hwqml	Reviewer panel 5 voices	1	5,296	\$0.5929	124.3
hwqml	Logical-fallacy report	1	2,217	\$0.5223	41.4
hwqml	Stage-6 CQE narrative	3,005	1,121	\$0.0761	31.9
hwqml total		3,008	10,180	\$1.6264	240.2
qcnn	Deep-research review	1	1,328	\$0.4381	37.5
qcnn	Reviewer panel 5 voices	1	4,932	\$0.4547	121.7
qcnn	Logical-fallacy report	1	2,494	\$0.3905	44.5
qcnn	Stage-6 CQE narrative	3,005	1,142	\$0.0758	29.7
qcnn total		3,008	9,896	\$1.3591	233.4
majorana2	Deep-research review	1	1,451	\$0.2954	39.8
majorana2	Reviewer panel 5 voices	1	5,072	\$0.3114	121.7
majorana2	Logical-fallacy report	1	3,112	\$0.2598	62.7
majorana2	Argument-structure audit	1	3,083	\$0.2662	73.5
majorana2	Numeric-claim registry (deterministic)	-	-	(est.)	0.0
majorana2	Disclosure audit	1	1,781	\$0.2270	33.7
majorana2	Anchored revision plan	2	2,676	\$0.3423	56.1
majorana2	Stage-6 CQE narrative	3,005	1,089	\$0.0763	28.3
majorana2 total		3,012	18,264	\$1.7784	415.8
Grand total		9,038	58,519	\$7.6682	1357.2

3 Verdict summary

Paper	Panel score	EIC verdict	CQE composite
lcutrotter	6.67/10	minor-revisions	23/100
flowvqe	6.0/10	major-revisions	23/100
hwqml	5.5/10	major-revisions	23/100
qcnn	6.0/10	major-revisions	23/100
majorana2	6.33/10	major-revisions	23/100

Panel score is the mean of the five voices’ *Verdict*: $N/10$ scores; the EIC verdict comes from the vote table; the CQE composite is the geometric mean of the six process-quality dimensions.

4 lcutrotter — *Simple and high-precision Hamiltonian simulation by compensating Trotter error with linear combination of unitary operations*

Pei Zeng, Jinzhao Sun, Liang Jiang, Qi Zhao [arXiv:2212.04566](https://arxiv.org/abs/2212.04566) PRX Quantum 6, 010359 (2025)

4.1 lcutrotter — Deep-research review

Model: claude-opus-4-5-20251101 · 1,336 output tokens · 36.5s

I'll analyze this paper systematically against the audit-and-falsify checklist.

4.1.1 1. One-Paragraph Summary

This paper proposes composite Hamiltonian simulation algorithms that combine Trotter formulas with Linear Combination of Unitaries (LCU) methods to achieve advantages of both approaches. The authors present two main algorithms: Paired Taylor-Series Compensation (PTSC), which achieves exponentially improved accuracy scaling over bare Trotter methods for generic Hamiltonians, and Nested-Commutator Compensation (NCC), which maintains near-linear system-size dependence for lattice Hamiltonians while quadratically improving accuracy. The key claim is that by adding few gates after a K-th order Trotter formula using LCU to compensate Trotter error, they achieve better time scaling $(1+1/(2K+1))$ instead of $1+1/K$ and dramatically improved accuracy—claimed to be 2 orders of magnitude better than fourth-order Trotter for generic Hamiltonians and 3-4 orders of magnitude higher accuracy for lattice systems at equivalent gate costs.

4.1.2 2. Audit-and-Falsify Checklist

Item	Status	Evidence
Augmented baseline catalog	PARTIAL	The paper compares against Kth-order Trotter [8,13,20] and post-Trotter methods [25,27], which are current methods, but Table I shows only asymptotic scalings without empirical head-to-head comparisons against specific recent implementations.
Strict-domination comparator	FAIL	Claims like “2 orders of magnitude smaller” and “3 to 4 orders of magnitude higher accuracy” (Sec. II.E, Fig. 8) are made at displayed precision without specifying calibrated tolerances (<code>_abs</code> , <code>_rel</code>) or error bars on these ratios.
Recompute-from-raw	PARTIAL	Fig. 8(a,b,c) show gate count comparisons, but there are no explicit tables of raw numerical values from which the displayed ratios could be independently verified; the comparison method for fourth-order Trotter is cited to Ref. [12,20] but intermediate values are not shown.
Wilson 95% CIs	NOT-APPLICABLE	This is a theoretical/analytical paper without sampling-based empirical results that would require binomial confidence intervals.
Cross-LLM falsifiability	NOT-APPLICABLE	No LLM-in-the-loop methodology was used in this work.
Honest negatives	FAIL	The paper does not include a Failure Modes section; there is no discussion of scenarios where the method underperforms, fails, or has limitations beyond general asymptotic regime requirements (e.g., what happens when x is not small).

Item	Status	Evidence
Simulator precision floor	PARTIAL	The paper is primarily analytical with asymptotic bounds; Fig. 8 shows gate count estimates but does not specify whether any numerical verification was performed at float64 vs complex64 precision.
Auditable claims	FAIL	No re-runnable script (e.g., <code>audit_claims.py</code>) or JSON artifacts are provided; the numerical claims in Fig. 8 lack accompanying raw data files or reproducible code.

4.1.3 3. Overall Assessment

This paper presents mathematically rigorous asymptotic complexity analysis with clear theoretical contributions. The core analytical results (Theorems 1 and 2, Propositions 3-7) appear sound, with proper derivations using established techniques (BCH formula, Taylor series, Euler’s formula for Pauli operators). However, from a research rigor standpoint, the paper has significant gaps: (1) the headline quantitative claims (“2 orders of magnitude,” “3-4 orders of magnitude”) lack precise calibration and raw data backing; (2) there is no discussion of failure cases or regimes where the method may not be advantageous; (3) the gate count comparisons in Fig. 8 rely on analytical bounds from prior work without direct numerical validation or error analysis; and (4) no reproducibility artifacts are provided. The paper would survive a theoretical review focused on mathematical correctness but would face challenges under a strict empirical audit demanding reproducible quantitative claims.

Research Rigor Score: 6/10

4.1.4 4. Three Highest-Leverage Improvements

- 1. Add explicit raw data tables and reproducibility artifacts:** Create supplementary material with (a) a table of all numerical values underlying Fig. 8, (b) the exact formulas/code used to compute gate counts for each method, and (c) a script that regenerates all figures from these raw values. This would address both “Recompute-from-raw” and “Auditable claims.”
- 2. Include a Failure Modes section:** Add explicit discussion of regimes where the method underperforms—e.g., when x approaches $1/(2)$, when the truncation order sc must be impractically large, or when the t sampling overhead from random-sampling LCU negates gate count advantages. Quantify the crossover points where standard Trotter becomes preferable.
- 3. Specify calibrated tolerances for comparative claims:** Replace vague claims like “2 orders of magnitude” with precise statements such as “at $= 10$ and $n = 20$, PTSC requires $1.2 \text{ } \mathbb{E} 10$ gates versus $1.8 \text{ } \mathbb{E} 10$ for fourth-order Trotter (ratio: $150 \mathbb{E}$, computed using bounds from [12] at the same).” Include the tolerance regime (`_abs` or `_rel`) and explicit formulas used for each method’s gate count.

4.2 lcutrotter — Reviewer panel 5 voices

Model: claude-haiku-4-5-20251001 · 5,293 output tokens · 122.0s

Peer Review Panel: PRX Quantum Submission

4.2.1 Paper: “Simple and high-precision Hamiltonian simulation by compensating Trotter error with linear combination of unitary operations”

4.2.2 Voice 1 — Reviewer 1 (Physics correctness)

The manuscript presents a compelling theoretical framework for combining Trotter formulas with linear combination of unitaries (LCU) to achieve improved Hamiltonian simulation performance. From a physics correctness standpoint, the core mathematical construction appears sound. The authors correctly identify that the K th-order Trotter remainder $V_K(x) = U(x)S_K(x)^\dagger$ satisfies the order condition $V_K(x) = I + O(x^{K+1})$, and their

subsequent Taylor expansion and pairing strategy exploits this property appropriately. The use of Euler’s formula (Eq. 7) to convert anti-Hermitian leading-order terms into Pauli rotations with suppressed 1-norm is mathematically elegant and correctly executed.

However, I have concerns regarding the treatment of specific physical Hamiltonians beyond the abstract lattice model. While the authors claim their method applies to “quantum chemistry Hamiltonians with large L ,” the paper provides no concrete analysis for molecular systems where the Hamiltonian structure differs significantly from the nearest-neighbor lattice models analyzed in detail. The nested-commutator bounds in Proposition 7 rely critically on the locality structure—specifically that $[H_{j,j+1}, H_{k,k+1}] = 0$ when $|j - k| > 1$. For electronic structure Hamiltonians in second quantization, the commutator structure is far more complex due to the non-local Coulomb integrals. The claim that “0th-order PTSC is particularly useful for quantum chemistry” (page 5) requires explicit verification beyond the L -dependence argument. The gate complexity $O(\lambda t)^2$ for 0th-order PTSC may be competitive for small t , but quantum chemistry simulations often require $t \sim 10^3$ for phase estimation, where the quadratic scaling becomes prohibitive.

The treatment of numerical precision is adequate for the theoretical framework but incomplete for practical implementation. The random-sampling implementation relies on Proposition 1, which bounds the estimation error as $|\langle O \rangle_V| \leq \|O\|(3\epsilon + \epsilon_n)$ with sample complexity $N = 2\mu^4 \ln(2/\delta)/\epsilon_n^2$. The μ^4 prefactor when $\mu = 2$ implies a 16CE overhead compared to standard sampling—this is stated but its implications for practical circuits deserve more attention. Furthermore, the truncation order s_c in Eq. (48) and (76) introduces systematic bias that depends on λx and the specific Hamiltonian; the authors provide asymptotic bounds but no finite-precision error analysis for realistic parameter regimes.

The comparison with fourth-order Trotter in Fig. 8 is physically meaningful, but I note the comparison uses analytical bounds from Ref. [12] rather than empirical tight bounds. For the Heisenberg model, tighter commutator-aware bounds exist (Proposition M.1 in Ref. [20]), which the authors do use for the NCC comparison. The asymmetry in bound tightness between the PTSC and Trotter comparisons may overstate the PTSC advantage. Additionally, the “2 orders of magnitude” improvement claim for PTSC (Abstract) relies on comparing against analytical fourth-order Trotter bounds, which are known to be loose by factors of 10-100CE in practice.

Verdict: 7/10

Recommendation: minor-revisions

Questions for Authors: 1. Can you provide explicit nested-commutator bounds for a molecular Hamiltonian (e.g., H in STO-3G basis) to validate the claim that PTSC is “particularly useful for quantum chemistry”? 2. How does the systematic truncation error at finite s_c compare to the statistical sampling error for realistic parameter choices? 3. Would tighter empirical Trotter bounds (rather than analytical bounds from Ref. [12]) change the claimed improvement factors in Fig. 8(a,b)? 4. What is the expected overhead for implementing the random sampling procedure on a fault-tolerant quantum computer with mid-circuit measurement and reset?

4.2.3 Voice 2 — Reviewer 2 (Algorithmic novelty)

The central contribution of this manuscript—compensating Trotter error with LCU formulas through order-pairing techniques—represents a genuine algorithmic innovation in the Hamiltonian simulation literature. The key insight that anti-Hermitian leading-order Trotter remainder terms can be paired with the identity using Euler’s formula to achieve 1-norm suppression from $1 + O((\lambda x)^{K+1})$ to $1 + O((\lambda x)^{2K+2})$ is novel and non-obvious. This effectively doubles the effective order of accuracy while maintaining the implementation simplicity of lower-order Trotter formulas.

Comparing against recent literature, this work should be contextualized against several relevant papers from 2023-2025. Hagan and Wiebe (Quantum 7, 1181, 2023) explored composite methods but did not achieve the order-pairing structure presented here. Cho, Berry, and Hsieh (Phys. Rev. A 109, 062431, 2024) developed randomized compensation techniques for Trotter errors, sharing conceptual similarities with the random-sampling implementation in Section II.D, but their approach does not achieve the commutator scaling that the NCC algorithm provides. The very recent work by Zhao et al. (Phys. Rev. Lett. 129, 270502, 2022) on time-dependent Hamiltonian simulation uses different techniques entirely. The authors correctly cite these works and distinguish their contributions, though the comparison with Ref. [47] (Cho et al.) deserves more explicit technical differentiation given the methodological overlap in using randomization to compensate Trotter errors.

The claimed complexity improvements in Table I represent legitimate Pareto improvements along specific dimensions. The PTSC algorithms achieve $\tilde{O}(\log(1/\epsilon))$ accuracy dependence (matching post-Trotter methods) while maintaining the $O(n^{1+1/(2K+1)})$ system-size scaling that improves upon the $O(n^2)$ of standard LCU/QSP methods for lattice Hamiltonians. The NCC algorithms achieve $O(\epsilon^{-1/(2K+1)})$ accuracy scaling with nearly-optimal $O(n^{1+2/(2K+1)})$ system-size dependence. These are not claimed as strict dominations across all dimensions—the authors honestly acknowledge trade-offs (e.g., PTSC has worse system-size dependence than Trotter for short times). The improvement ratios in Fig. 8(c) showing “3-4 orders of magnitude higher accuracy” are recomputable from the analytical bounds in Section V.

However, I question whether the claimed novelty fully accounts for the relationship with existing qDRIFT-type algorithms. The random-sampling implementation (Fig. 2) is essentially a structured variant of qDRIFT applied to the Trotter remainder. While the authors cite Refs. [33-35] appropriately, the distinction between their approach and Campbell’s qDRIFT (Phys. Rev. Lett. 123, 2019) applied with Trotter pre-processing deserves explicit analysis. Specifically, what prevents one from running first-order Trotter followed by qDRIFT on the multiplicative error $V_K(x)$? The pairing technique provides the novel element, but the random-sampling infrastructure is inherited.

The algorithmic contribution is substantive but not transformative. The complexity improvements are incremental (polynomial factors) rather than asymptotic class changes. For lattice Hamiltonians, the practically relevant improvement is reducing gate counts by constant factors (the “2-4 orders of magnitude” claims) rather than improving scaling exponents from $O(n^{1.25})$ to $O(n^{1.2})$. The paper would be strengthened by explicit resource estimates for a specific target application (e.g., simulating a 100-qubit Heisenberg chain to chemical accuracy) comparing total T-gate counts across all methods.

Verdict: 7/10

Recommendation: minor-revisions

Questions for Authors: 1. How does your method compare to applying qDRIFT directly to the Trotter remainder $V_K(x)$ without the pairing technique? What is the quantitative advantage of pairing? 2. Can you provide an explicit resource comparison (T-gate counts, circuit depth) for a specific application benchmark such as simulating the Fermi-Hubbard model at half-filling? 3. The improvement from $O(t^{1+1/K})$ to $O(t^{1+1/(2K+1)})$ is less significant at high orders—is there an optimal K for practical implementations? 4. For the coherent implementation (Appendix H), how do the ancilla qubit requirements compare to standard QSP implementations?

4.2.4 Voice 3 — Reviewer 3 (Empirical evidence)

The empirical evidence presented in this manuscript is primarily analytical rather than numerical, which is appropriate for the theoretical nature of the contribution but raises questions about practical validation. The main numerical results appear in Fig. 8, which compares gate counts based on analytical bounds rather than explicit circuit compilation. While this approach is standard in complexity-theoretic Hamiltonian simulation papers, it limits the ability to verify the claimed improvements in practice.

The gate counting methodology in Section II.E and Fig. 8 requires scrutiny. The authors state they “compile their quantum circuits to CNOT gates, single-qubit Clifford gates, and single-qubit Z-axis rotation gates $R_z(\theta)$ ” and count R_z gates as the resource metric. However, the actual circuit structure for the random-sampling implementation differs from standard Trotter circuits. The controlled-Pauli and controlled-Pauli-rotation gates in Fig. 11 require decomposition that depends on the sampled Pauli weight, which is a random variable. The claimed gate counts should therefore be expected values over the sampling distribution, but the authors do not explicitly compute these expectations—they bound the worst-case Pauli weight by $O(s_c)$. For PTSC with $s_c \sim \log(1/\epsilon)/\log \log(1/\epsilon)$, this introduces logarithmic factors that may not be negligible. A proper empirical validation would sample many random instances and report the distribution of gate counts.

The comparison with fourth-order Trotter uses bounds from Ref. [12] (analytical) and Ref. [20] (commutator-aware), but these represent different tightness levels. The PTSC comparison in Fig. 8(a,b) uses Ref. [12], while the NCC comparison in Fig. 8(c) uses Ref. [20]. This asymmetry is disclosed but complicates interpretation. An honest comparison would use the tightest available bounds for all methods. Furthermore, the y-axis label “Rz Gate” conflates different gate definitions: for Trotter, these are deterministic R_z gates in the circuit; for Trotter-LCU, these include random Pauli rotations whose angles depend on the LCU formula parameters. The resource overhead of computing these angles classically is not accounted for.

The paper lacks statistical analysis appropriate for numerical claims. The “2 orders of magnitude” and “3-4 orders of magnitude” improvement claims are point estimates from analytical formulas, not sample statistics with confidence intervals. While this is common in theoretical papers, the claims would be strengthened by: (1) implementing the sampling procedure in Algorithm 1 and verifying the claimed distribution over Pauli operators; (2) running explicit simulations of small systems (e.g., 4-6 qubits) to verify that the LCU formulas achieve the claimed approximation errors; (3) reporting variance in gate counts across different random samples. The absence of any failure mode analysis or honest-negatives section is notable—the paper does not discuss scenarios where Trotter-LCU might underperform, such as when λt is small or when the Hamiltonian has non-local structure.

The Heisenberg model numerical example in Fig. 8(c) and Algorithm 1 provides the most concrete validation. Algorithm 1 is explicit enough to be reproduced, and the parameter choices ($\theta := \tan^{-1}(16nx^2(1+24x))$) can be verified against the analytical formulas. However, the paper does not report any actual execution of this algorithm—it remains a specification rather than an implementation. An accompanying code repository with scripts to regenerate Fig. 8 would substantially strengthen the empirical contribution.

Verdict: 6/10

Recommendation: major-revisions

Questions for Authors: 1. Can you provide code or pseudocode that reproduces the gate counts in Fig. 8 from the analytical formulas? 2. What is the expected Pauli weight distribution for the sampled operators in Algorithm 1, and how does this affect the average gate count? 3. Have you implemented the sampling procedure and verified the claimed LCU approximation errors on small systems? 4. Under what conditions (e.g., short times, highly non-local Hamiltonians) does your method underperform compared to standard Trotter or QSP?

4.2.5 Voice 4 — Devil’s Advocate

This paper should be rejected or require major revisions due to several fundamental issues that the other reviewers have treated too charitably.

The claimed improvements are primarily artifacts of comparison methodology, not genuine algorithmic advances. The “2 orders of magnitude” improvement in Fig. 8(a,b) compares PTSC against the *analytical* fourth-order Trotter bound from Ref. [12], which is known to be extremely loose. Childs et al. (Ref. [20]) explicitly demonstrate that commutator-aware bounds can be orders of magnitude tighter. When the authors switch to commutator-aware bounds for the NCC comparison in Fig. 8(c), the improvement shrinks to “3-4 orders of magnitude *in accuracy*”—which means achieving $\epsilon = 10^{-6}$ instead of 10^{-3} with the same gate count. But this comparison uses second-order NCC against fourth-order Trotter; a fair comparison would use fourth-order NCC (if the analysis were tractable, which the authors admit it is not: “We leave precise higher-order NCC gate count analysis for future study”). The paper’s headline claims are thus based on asymmetric comparisons that systematically favor the new method.

The random-sampling implementation has hidden costs that undermine the complexity claims. Proposition 1 states that random-sampling LCU requires $N = O(\mu^4/\epsilon_n^2)$ samples. With $\mu = 2$ (as used in the numerical comparisons), this is a 16E overhead that the authors acknowledge but downplay. More critically, the sampling procedure in Fig. 5 and Algorithm 2 requires *classical* computation of multinomial distributions, Pauli products, and angle parameters that scale with the system size and truncation order. The claimed gate complexity $O((t)^{1+1/(2K+1)}(\kappa_K L + \log(1/\epsilon)/\log \log(1/\epsilon)))$ (Theorem 1) counts only quantum gates, ignoring the classical preprocessing cost. For the NCC algorithm, the space cost is $O(K\kappa)$ and time cost is $O(K(\log \kappa + \log n))$ per sample (Appendix D), but with $\kappa = 2 \times 5^{K/2-1}$ for $K = 2k$, this grows exponentially in K . The fourth-order NCC algorithm would have $\kappa = 10$, making the classical sampling overhead non-negligible for large systems.

The nested-commutator compensation lacks practical implementation details. The NCC algorithm requires computing the explicit nested-commutator expansion of the Trotter remainder, which involves exponentially many terms (see Eq. (C17)). The “padding” technique in Section II.D and Fig. 6 introduces virtual ancilla qubits and zero-valued commutators to achieve uniform sampling, but the overhead of this padding is not quantified. For the first-order Heisenberg example (Algorithm 1), the structure is simple, but extending to higher orders or more complex Hamiltonians requires case-by-case analysis that the authors have not completed. The claim that NCC is “easy to implement” (Table I) is misleading for practical implementations.

The paper lacks honest acknowledgment of failure modes. The method *requires* $\lambda x < 1/(2\lambda)$ (Proposition 3) for the Taylor expansion to converge, which constrains the minimum segment number $\nu \geq 2\lambda t$.

For long-time simulations ($t \sim 10^3$) of large systems ($\lambda \sim n$), this requires $\nu \sim 2000n$ segments, each involving random sampling. The variance of random-sampling estimators grows with the number of sequential samples, introducing error accumulation that is not analyzed. The paper also does not discuss the practical challenge of implementing mid-circuit measurement and reset (Fig. 2(c)) on current fault-tolerant architectures, which may introduce significant overhead.

Recommendation: major-revisions

The paper contains a valid theoretical contribution (the order-pairing technique) buried under overstated claims and incomplete analysis. To merit publication in PRX Quantum, the authors must: (1) provide symmetric comparisons using the tightest available bounds for all methods; (2) quantify the classical preprocessing and sampling overhead; (3) implement the algorithms on small systems to validate the theoretical claims; (4) honestly discuss failure modes and parameter regimes where the method underperforms.

4.2.6 Voice 5 — Editor-in-Chief synthesis

Having reviewed all four assessments, I find substantive merit in the theoretical contribution alongside legitimate concerns about the empirical validation and comparison methodology. The Devil’s Advocate raises valid points that require response, though some criticisms are more central than others.

The core theoretical contribution—using Euler’s formula to pair anti-Hermitian Trotter remainder terms with the identity, thereby doubling the effective order of 1-norm suppression—is novel and mathematically sound. Reviewers 1 and 2 agree on this point. The resulting complexity improvements in Table I represent genuine (if incremental) advances: PTSC achieves logarithmic accuracy dependence while maintaining sub-quadratic system-size scaling for structured Hamiltonians, and NCC achieves better-than-Trotter accuracy scaling with commutator-aware system-size bounds. These are valuable contributions to the Hamiltonian simulation toolkit.

However, the Devil’s Advocate correctly identifies that the comparison methodology is asymmetric and potentially misleading. The “2 orders of magnitude” headline claim compares PTSC against loose analytical Trotter bounds, while the NCC comparison uses tighter commutator-aware bounds. This inconsistency must be addressed before publication. Additionally, Reviewer 3’s concern about the lack of numerical implementation is well-founded—for a paper claiming practical improvements of multiple orders of magnitude, some empirical validation beyond analytical bounds is expected, even for a theory-focused venue like PRX Quantum.

The classical preprocessing overhead raised by the Devil’s Advocate (exponential in K for NCC due to $\kappa = 2 \times 5^{K/2-1}$) is a legitimate concern for high-order implementations, but the paper primarily advocates for low-order methods (first or second order) where $\kappa \leq 2$. The authors should clarify this limitation explicitly. The 16E sampling overhead (μ^4 with $\mu = 2$) is disclosed and is the cost of the random-sampling implementation; the coherent implementation in Appendix H avoids this overhead at the cost of more complex circuits.

Regarding Reviewer 1’s questions about quantum chemistry applications: the paper should either provide explicit analysis for molecular Hamiltonians or remove the claim that PTSC is “particularly useful” for this setting. The lattice Hamiltonian analysis is thorough; extending claims beyond this domain requires comparable rigor.

Final Verdict: minor-revisions

The paper presents a valid and novel algorithmic contribution suitable for PRX Quantum, but requires revisions to address methodological concerns before acceptance.

Must-fix items before resubmission (ordered by severity):

1. **Symmetric comparison methodology:** Regenerate Fig. 8(a,b) using commutator-aware Trotter bounds (Ref. [20] methodology) rather than loose analytical bounds from Ref. [12], or provide explicit justification for why the looser bounds are appropriate.
2. **Quantify classical overhead:** Add a subsection or paragraph explicitly stating the classical preprocessing cost (space and time) for the sampling procedures, particularly noting how κ scales with Trotter order K and the implications for practical implementations.
3. **Remove or substantiate quantum chemistry claims:** Either provide explicit nested-commutator bounds for a molecular Hamiltonian or remove the claim that PTSC is “particularly useful for quantum chemistry” (Section II.B).

4. **Add failure modes discussion:** Include a paragraph discussing parameter regimes where Trotter-LCU underperforms (e.g., short simulation times, highly non-local Hamiltonians, small accuracy requirements where the sampling overhead dominates).
5. **Provide reproducibility resources:** Include code or detailed pseudocode sufficient to reproduce the gate counts in Fig. 8, or commit to providing a code repository upon acceptance.
6. **Clarify comparison with qDRIFT:** Add explicit analysis distinguishing the pairing technique from simply applying qDRIFT to the Trotter remainder, including quantitative comparison if possible.

4.2.7 Vote table

Voice	Recommendation	Confidence 1-10
Reviewer 1	minor-revisions	7
Reviewer 2	minor-revisions	7
Reviewer 3	major-revisions	6
Devil's Advocate	major-revisions	8
Editor-in-Chief	minor-revisions	7

4.3 lcutrotter — Logical-fallacy report

Model: claude-opus-4-5-20251101 · 2,348 output tokens · 45.4 s

Finding 1

- **Fallacy:** cherry-picked-baseline
- **Severity:** medium
- **Location:** Section I (Introduction), paragraph 3; Table I
- **Evidence:** “Trotter methods are recently rigorously shown to enjoy commutator scaling [13, 20]. . . . Consequently, for instance, for n-qubit lattice Hamiltonians, their gate complexities are $O(n^{\check{s}})$, which is worse than those in Trotter algorithms $O(n^{1+o(1)})$.”
- **Why it’s the fallacy:** The paper frames the comparison against “post-Trotter” algorithms (Taylor series, QSP) as having $O(n^{\check{s}})$ system-size scaling for lattice Hamiltonians, but this characterization applies only when these methods do not exploit commutator structure. Recent work has shown that post-Trotter methods can also achieve better scaling when commutator bounds are incorporated. The paper selectively highlights the worst-case scaling of competitors while showcasing the best-case scaling of their own method.
- **Suggested fix:** Acknowledge that post-Trotter methods can also be improved with commutator-aware implementations, and clarify that the $O(n^{\check{s}})$ scaling applies to the standard implementations without such optimizations. Add a sentence such as: “We note that post-Trotter methods could potentially be improved by incorporating commutator structure, though such implementations are not yet standard.”

Finding 2

- **Fallacy:** asymptotic-only-claim
- **Severity:** medium
- **Location:** Section II.E (Performance comparison), paragraph 1; Theorem 1 and Theorem 2
- **Evidence:** “From Table I, we observe that. . . the NCC gate counts show improved system-size dependence.” and “the gate complexity of random-sampling Kth-order Trotter-LCU algorithm. . . is $O(n^{1+2/(2K+1)} t^{1+1/(2K+1)})$.”

- **Why it’s the fallacy:** The paper makes strong asymptotic claims about improved scaling (e.g., time dependence improving from $t^{\{1+1/K\}}$ to $t^{\{1+1/(2K+1)\}}$), but the numerical demonstrations in Fig. 8 are limited to relatively modest system sizes (n up to 100) and short to moderate evolution times. The crossover points where the asymptotic advantages manifest are not clearly demonstrated, and constant factors hidden in the $O(\hat{u})$ notation could dominate at practical scales.
 - **Suggested fix:** Include explicit analysis of crossover points where the proposed algorithms outperform baselines. Add a statement such as: “The asymptotic improvements become dominant when [specific conditions on n , t , are met]; for smaller instances, constant factors may favor simpler methods.”
-

Finding 3

- **Fallacy:** conflated-regimes
 - **Severity:** medium
 - **Location:** Section II.E, Figure 8(a,b) and surrounding discussion
 - **Evidence:** “For a generic Hamiltonian, the estimated gate counts of the first algorithm can be 2 orders of magnitude smaller than the best analytical bound of fourth-order Trotter formula.”
 - **Why it’s the fallacy:** The 2-local Hamiltonian $H = \sum_{\{i,j\}} X_i X_j + \sum_i Z_i$ used for the “generic” comparison in Fig. 8(a,b) has all-to-all connectivity and specific structure. Extrapolating claims of “2 orders of magnitude” improvement to truly generic L -sparse Hamiltonians (including, e.g., quantum chemistry Hamiltonians with irregular structure) may not hold. The paper conflates performance on this specific benchmark with general applicability.
 - **Suggested fix:** Qualify the claims by specifying that the 2-order improvement is demonstrated for the specific 2-local model with uniform couplings. Add: “These results are specific to the 2-local model studied; performance on other Hamiltonians with different structure may vary.”
-

Finding 4

- **Fallacy:** hasty-generalization
 - **Severity:** medium
 - **Location:** Section VI (Conclusion), paragraph 1
 - **Evidence:** “We study the Hamiltonian simulation algorithms based on the composition of Trotter and LCU algorithms. In both theoretical and numerical studies, we show that the 0th-order paired Taylor-series compensation (PTSC) algorithm, 2kth-order PTSC algorithm and the 2kth-order nested-commutator compensation (NCC) algorithm enjoy different advantages and will be useful in different scenarios.”
 - **Why it’s the fallacy:** The paper generalizes from a limited set of numerical experiments (primarily Heisenberg model and 2-local Hamiltonian) to broad claims about when each algorithm “will be useful.” No systematic benchmarking across diverse Hamiltonian families (e.g., molecular Hamiltonians, power-law interactions, disordered systems) is provided to support these scenario-based recommendations.
 - **Suggested fix:** Temper the conclusion by noting the limited scope of numerical validation: “Based on our analysis of the Heisenberg model and 2-local Hamiltonians, we conjecture that... Further benchmarking on other Hamiltonian classes is needed to confirm these recommendations.”
-

Finding 5

- **Fallacy:** ad-hoc-precision-floor
- **Severity:** medium
- **Location:** Section II.E, Figure 8(c)
- **Evidence:** “Particularly, using the same gate number as the fourth-order Trotter, we are able to achieve a 3 to 4 orders of magnitudes higher accuracy .”

- **Why it’s the fallacy:** The claim of “3 to 4 orders of magnitude higher accuracy” compares algorithmic error bounds, not actual achieved accuracy in the presence of realistic noise sources (gate errors, decoherence, sampling variance from the random LCU implementation). The values ranging from 10^{-3} to 10^{-6} in Fig. 8(c) are below typical noise floors for near-term quantum devices, and even for fault-tolerant devices, such precision claims should account for the t^4 sampling overhead acknowledged in Proposition 1.
- **Suggested fix:** Clarify that the accuracy comparison is between analytical error bounds assuming perfect implementation, and acknowledge that achieving such precision requires accounting for sampling overhead: “These accuracy comparisons reflect analytical bounds; achieving $= 10^{-6}$ in practice requires $O(t^4/s)$ samples, which may be substantial.”

4.4 lcutrotter — Stage-6 CQE narrative

Model: claude-haiku-4-5-20251001 · 1,195 output tokens · 31.9s

Process Summary: QuantumNovelty Run Evaluation

4.4.1 Composite Verdict

The run achieved a **composite score of 23 out of 100**, calculated via geometric mean across six dimensions. Per standard interpretation scales, this places the work firmly in the “**Needs Substantial Work**” tier (scores below 30 indicate fundamental gaps in methodology or execution). A score of 23 signals that while some scaffolding exists, the run failed to produce the core artifacts necessary for a credible novelty claim. This is not a matter of polish—it reflects missing foundational components.

To be direct: a composite of 23 means the run cannot support any publishable or actionable conclusions in its current state. The geometric mean methodology is unforgiving by design—it penalizes runs that neglect entire dimensions rather than excelling in a few while ignoring others. Here, no dimension exceeded 40, and the lowest scored just 8. The geometric mean correctly surfaces that this run has systemic, not localized, problems.

4.4.2 Strongest Dimension: Communication (40)

The **Communication** dimension scored highest at 40, though this requires careful interpretation. Both probes—“logical fallacies absent” and “reviewer panel verdict”—scored 40, but the evidence reveals these scores derive from *absence of negative signal* rather than *presence of positive verification*. The “logical fallacies skill not run” and “no review_panel.md found” evidence strings indicate these probes weren’t executed at all. A score of 40 here essentially means “we didn’t detect problems because we didn’t look.”

This is a ceiling imposed by non-execution, not a floor established by quality. What this says about the run is revealing: the communication layer was deprioritized entirely. No reviewer simulation occurred. No logical consistency check ran. The “strongest” dimension is strongest only because incomplete execution produces ambiguous rather than definitively poor results. In a future run, actually executing these probes could easily *lower* this score if fallacies or reviewer objections surface.

4.4.3 Weakest Dimension: Novelty Rigour (8)

The **Novelty Rigour** dimension scored a critically low 8, making it the clear failure point of this run. This dimension is arguably the most important for a QuantumNovelty workflow—without rigorous novelty verification, the entire purpose of the run is compromised.

The probe-level breakdown is damning:

- “**augmented baseline catalog present**” scored 10, with evidence showing “baseline_catalog has 0 rows.” This means the run produced no baseline against which to compare results. Without a baseline catalog, there is no reference frame for claiming novelty—any “discovery” could be trivial rediscovery of known results.

- **“strict-domination comparator run”** scored 5, with “novelty_verdict.json not found.” The strict-domination comparator is the mechanism that determines whether a candidate solution genuinely dominates known solutions or merely interpolates between them. Its absence means no formal novelty adjudication occurred.

The specific stage that produced this failure was almost certainly the **baseline construction and comparison phase**. Either the baseline retrieval failed silently, or the run proceeded without waiting for baseline population. The downstream comparator couldn’t run because there was nothing to compare against. This is a pipeline ordering or dependency-management failure, not a methodological design flaw.

4.4.4 Three Highest-Leverage Improvements

1. Gate the Pipeline on Baseline Catalog Population The single highest-leverage fix is adding a hard gate that blocks downstream stages until `baseline_catalog` contains a minimum viable row count (suggest 50 known solutions for quantum chemistry problems of typical complexity). The “augmented baseline catalog present” probe should trigger a pipeline halt, not merely log a warning, when rows equal zero. This prevents the entire novelty evaluation apparatus from running vacuously.

2. Mandate Artifact Emission Before Scoring Multiple probes failed because expected artifacts simply don’t exist: `novelty_verdict.json`, `audit_claims.py`, `paper.tex`, `wilson_annotations.md`, `ablation_results.json`, `ratio_recompute.md`, `review_panel.md`. The Reproducibility dimension (20) and Methodological Rigour dimension (27) both suffer from missing files. Implement a pre-scoring checklist that verifies required artifacts exist and are non-empty. If critical files are absent, the run should fail loudly rather than produce a misleading low score.

3. Execute Cross-LLM Validation with At Least Two Vendors The Falsifiability dimension (30) was dragged down by “vendors used: []”—zero cross-LLM validation occurred. Single-model runs cannot distinguish genuine algorithmic discoveries from model-specific artifacts or memorization. Requiring at least two vendors (e.g., one Anthropic model plus one OpenAI or open-weights model) before a run is considered complete would substantially increase confidence that any findings generalize.

Summary: This run produced a scaffold without substance. The composite score of 23 accurately reflects a workflow that executed structurally but failed to populate the artifacts that give structure meaning. The path forward is mechanical: enforce baseline population, require artifact existence, and diversify model vendors. These are not creative problems—they are engineering discipline problems with straightforward solutions.

5 flowvqe — *Generative flow-based warm start of the variational quantum eigensolver*

Hang Zou, Martin Rahm, Anton Frisk Kockum, Simon Olsson [arXiv:2507.01726](https://arxiv.org/abs/2507.01726) npj Quantum Information (2025)

5.1 flowvqe — Deep-research review

Model: claude-haiku-4-5-20251001 · 1,318 output tokens · 38.0 s

5.1.1 1. One-paragraph summary of what the paper claims

The paper introduces Flow-VQE, a framework that uses conditional normalizing flows to generate high-quality variational parameters for the Variational Quantum Eigensolver (VQE). The authors claim that by embedding a generative model into the VQE optimization loop via preference-based training, Flow-VQE enables gradient-free optimization and provides a systematic approach for parameter transfer across related molecular systems. Through numerical simulations on H, HO, NH, and CH, they report that Flow-VQE achieves computational accuracy with fewer circuit evaluations than baseline optimizers (improvements ranging from modest to over two orders of magnitude), and when used for warm-starting, accelerates subsequent fine-tuning by up to 50-fold compared to Hartree-Fock initialization. The total training overhead is claimed to be no greater than optimizing five molecules by conventional methods.

5.1.2 2. Audit-and-falsify checklist

Criterion	Verdict	Evidence
Augmented baseline catalog	PARTIAL	The paper compares against GD, Adam, and QNSPSA, which are standard VQE optimizers, but does not benchmark against recent ML-based warm-start methods (e.g., meta-learning approaches [40-44], supervised learning [32,33], or other generative approaches [34,35]) that they cite in the introduction.
Strict-domination comparator	FAIL	Claims like “up to two orders of magnitude” and “50-fold acceleration” are stated without specifying calibrated tolerances (<code>_abs</code> , <code>_rel</code>); Figures 2 and 4 show bar plots without error bars or statistical uncertainty quantification on the circuit evaluation counts.
Recompute-from-raw	FAIL	No raw numerical tables are provided in the paper or supplementary material; ratios and improvement factors cannot be independently verified from tabulated source data (Table I provides post-training comparison but not the raw data underlying Figures 2-4).
Wilson 95% CIs	FAIL	No confidence intervals are reported on any results; particularly concerning for small-sample comparisons (e.g., 6 geometries for HO, 8 for H) and the batch size $B=2$ sampling regime where statistical fluctuations would be substantial.
Cross-LLM falsifiability	NOT-APPLICABLE	This paper does not use an LLM-in-the-loop method; the generative model is a normalizing flow trained via preference optimization, not a language model.

Criterion	Verdict	Evidence
Honest negatives	PARTIAL	The paper acknowledges limitations in Section VI (reduced diversity over training, potential loss of exploration, less precision than gradient-based methods in smooth landscapes), but does not include a dedicated Failure Modes section showing specific cases where Flow-VQE failed to improve over baselines or failed to converge.
Simulator precision floor	FAIL	No mention of float64 vs complex64 precision; all experiments use PennyLane state-vector simulations without specifying numerical precision, which is relevant given the “computational accuracy” threshold of 1.6×10^{-5} Hartree.
Auditable claims	FAIL	No <code>audit_claims.py</code> or equivalent script is provided; no mention of code/data availability in the paper body (required statements for npj Quantum Information are absent from this preprint version).

5.1.3 3. Overall assessment

This paper presents a creative and potentially impactful idea—using normalizing flows with preference-based training to warm-start VQE—but the empirical validation falls significantly short of the rigour expected for claims of “up to two orders of magnitude” improvement. The absence of error bars, confidence intervals, or statistical tests on performance metrics undermines confidence in the reported speedups. The baseline comparisons exclude the very ML-based methods the paper positions itself against in the introduction. No raw data tables or reproducibility scripts are provided, making independent verification impossible. The paper would likely not survive a strict reviewer-mode audit in its current form: the methodology is interesting but the evidence is presented at display precision without uncertainty quantification, and several required statements for the target journal (Code Availability, Data Availability) are missing from the preprint.

Research rigour score: 4/10

5.1.4 4. Three highest-leverage improvements

- Add statistical uncertainty to all quantitative claims:** Report Wilson 95% confidence intervals or bootstrap CIs for circuit-evaluation counts, especially given the small number of molecular geometries (6-8 per system). Run multiple random seeds for all methods and report mean \pm std across seeds. The “50-fold” and “two orders of magnitude” claims should be converted to ratios with confidence bounds (e.g., “27 \times [95% CI: 18-41]”).
- Benchmark against ML-based warm-start methods:** The paper cites meta-learning [40-44], supervised learning [32,33], and generative approaches [34,35] but only compares against traditional optimizers. Include at least one ML-based warm-start baseline (e.g., FLIP [42] or a supervised neural-network initializer) to demonstrate strict domination over current published methods, not just textbook baselines.
- Provide auditable artifacts and raw data:** Create an `audit_claims.py` (or equivalent) that derives every numerical claim (improvement ratios, circuit counts, energy errors) from on-disk JSON/CSV files. Include a Data Availability and Code Availability statement with a permanent repository link. Specify the numerical precision (float64/complex64) used in all simulations and provide the exact random seeds for reproducibility.

5.2 flowvqe — Reviewer panel 5 voices

Peer Review Panel: Flow-VQE Manuscript

5.2.1 Voice 1 — Reviewer 1 (Physics correctness)

The manuscript presents a methodologically sound approach to variational quantum eigensolver optimization through normalizing flows. The Hamiltonian construction follows standard procedures: Jordan-Wigner transformation for fermion-to-qubit mapping, active space selections that are physically reasonable ((4e,4o) for H, (6e,5o) for HO, (6e,6o) for NH and CH), and appropriate basis sets (STO-3G for the larger molecules, cc-pVDZ for H). The choice of Jordan-Wigner over Bravyi-Kitaev or parity mappings is not justified but is defensible given the modest qubit counts (5-12 qubits). The Z symmetry tapering applied to H is correctly implemented and reduces the qubit count from 8 to 5, which is standard practice.

The Hartree-Fock reference initialization strategy is well-motivated. The authors correctly note that HF typically recovers >99.5% of total electronic energy for small closed-shell molecules, making the remaining correlation energy the critical optimization target. The ansatz choices are appropriate: the hardware-efficient RY-linear ansatz for H and Givens-based singles and doubles (GSD) for the other molecules. The GSD ansatz preserves particle number and spin symmetry by construction, which is crucial for chemical accuracy. However, the manuscript does not discuss the expressibility limits of these ansätze in relation to the active spaces chosen—particularly whether the GSD ansatz with 54-117 parameters can capture the full correlation energy within the selected active spaces.

One concern is the absence of discussion regarding numerical precision. The simulations are performed via PennyLane state-vector simulation, but the manuscript does not specify whether complex128 or complex64 precision was used. For stretched geometries where strong correlation effects dominate (e.g., H at 2.6 Å or HO at 1.9 Å), numerical precision can significantly affect the computed energies, particularly when comparing against “exact diagonalization at the same level of theory.” The claimed computational accuracy threshold of 1.6E10 μ Hartree (\sim 1 kcal/mol) is chemically meaningful, but the authors should verify that numerical artifacts do not contaminate their comparisons at this precision level.

The treatment of stretched bond regions deserves additional scrutiny. The authors acknowledge that “energy errors increase in stretched bond-length regions due to strong correlation effects and the limited expressivity of the employed ansätze.” This is physically correct, but the manuscript would benefit from quantifying the multireference character (e.g., via T1 diagnostic or natural orbital occupation numbers) to distinguish between ansatz limitations and Flow-VQE performance. Additionally, the equilibrium geometries and bond stretching ranges are reasonable for benchmarking, but the ammonia inversion coordinate and benzene C-H stretch represent different classes of nuclear motion that test different aspects of the PES—this diversity is a strength, but the physical rationale for these choices could be made more explicit.

Questions for Authors: 1. What numerical precision (complex64 vs complex128) was used in the state-vector simulations, and have you verified that precision artifacts do not affect comparisons at the 1.6E10 μ Hartree threshold? 2. Can you provide multireference diagnostics (T1 amplitudes or natural orbital occupations) for the stretched geometries to distinguish ansatz expressibility limitations from optimization performance? 3. Why was Jordan-Wigner chosen over Bravyi-Kitaev, given that BK can reduce circuit depth for certain ansätze? 4. For the GSD ansatz, have you verified that 54-117 parameters are sufficient to reach the FCI limit within your active spaces?

Verdict: 7/10 — Minor Revisions

5.2.2 Voice 2 — Reviewer 2 (Algorithmic novelty)

The core algorithmic contribution—using conditional normalizing flows trained via preference-based optimization to generate VQE parameters—represents a meaningful advance over prior work, though the novelty claims require careful contextualization. The manuscript correctly identifies key limitations of existing approaches: gradient-based methods incur $O(d)$ overhead per iteration, gradient-free methods scale poorly with dimensionality, and conventional parameter transfer relies on geometric proximity heuristics. Flow-VQE addresses these through a learned conditional prior that can generalize across chemical space.

Comparing against recent literature (2023-2025), several closely related works warrant discussion. Rudolph et al. (Nat. Commun. 2023, Ref. [25]) demonstrated tensor-network pretraining for parameterized quantum

circuits, achieving similar goals of informed initialization. The manuscript does cite this but does not provide direct numerical comparison. More critically, Nakaji et al. (arXiv 2401.09253, Ref. [80]) introduced the Generative Quantum Eigensolver (GQE), which generates entire quantum circuits rather than just parameters—a more ambitious scope that subsumes Flow-VQE’s approach. The authors acknowledge this in Section VI as future work but should more directly address how Flow-VQE relates to GQE’s published results. Additionally, Chang et al. (arXiv 2505.10842, Ref. [44]) propose LSTM-based parameter prediction specifically for VQE, published after the apparent submission of this work but representing convergent thinking that affects novelty assessment.

The preference-based optimization scheme (Section III.C) is the most novel technical contribution. Drawing from RLHF methods in language models (DPO, Ref. [68]), the authors replace high-variance policy gradients with pairwise comparisons and elite buffer maintenance. This is clever and well-motivated: the observation that “chemically meaningful energy differences translate to extremely weak learning signals” (Section III.B.1) is correct and explains why vanilla REINFORCE would struggle. However, the connection to established preference learning theory is somewhat superficial—the authors do not discuss the implicit reward model induced by their approach or how it relates to Bradley-Terry preferences.

Regarding claimed performance ratios, the headline numbers (“up to two orders of magnitude fewer circuit evaluations,” “50-fold acceleration”) require scrutiny. Examining Figure 2, the two-orders-of-magnitude claim appears to derive from comparing Flow-VQE-S against gradient descent at specific bond lengths (e.g., HO at 1.8 Å shows $\sim 10\times$ improvement). However, the appropriate baseline comparison should be against Adam with properly tuned learning rates, where improvements are 2-5 \times according to the authors’ own text. The 50-fold warm-start acceleration (Table I) occurs at learning rate =0.001, which is unrealistically conservative for standard VQE—practitioners typically use [0.01, 0.1]. At =0.02, the improvement is $\sim 27\times$ for HO and $\sim 11\times$ for H, which are still impressive but more modest. These nuances should be foregrounded rather than buried.

Questions for Authors: 1. Can you provide direct numerical comparison against tensor-network pretraining (Rudolph et al.) and/or GQE (Nakaji et al.) on overlapping benchmark molecules? 2. What implicit reward model does your preference-based training induce, and how does it relate to Bradley-Terry or Plackett-Luce models? 3. Why were baseline comparisons performed at =0.02 rather than grid-searching for optimal baseline learning rates, given that your headline improvements depend on baseline performance?

Verdict: 6/10 — Major Revisions

5.2.3 Voice 3 — Reviewer 3 (Empirical evidence)

The experimental methodology presents several concerns regarding statistical rigor, reproducibility, and completeness of ablations. While the numerical results are promising, the empirical evidence does not meet the standards expected for a high-profile computational quantum chemistry publication.

The most significant gap is the absence of uncertainty quantification. All reported circuit evaluation counts and energy errors are point estimates without confidence intervals or multi-seed variance. Given the stochastic nature of both the normalizing flow sampling and the preference-based training, results should be reported over multiple random seeds (n_5 is typical). For example, the claim that “Flow-VQE-S achieves approximately a two- to five-fold reduction” over Adam lacks error bars—does this range reflect variation across molecules, or could it also reflect run-to-run variance that would widen the comparison? Table I reports final energy errors to 4 significant figures (e.g., 5.932 $\times 10^{-4}$ Hartree), but without standard deviations, these precision claims are unverifiable.

The ablation study is incomplete. The authors introduce several design choices—Gaussianization flows versus alternative architectures, 7-20 flow layers, buffer size $M=2$, batch size $B=2$, Gaussian noise regularization ($\xi=0.001$), and linear Hamiltonian embeddings—but do not systematically ablate their contributions. Section VI acknowledges that “Flow-VQE introduces some additional hyperparameters... which may require more empirical tuning,” but the reader is left without guidance on which choices are critical. Particularly concerning is the elite buffer size $M=2$: with only two samples retained per configuration, the training could be highly sensitive to early lucky draws. An ablation varying $M\{1, 2, 5, 10\}$ would clarify whether the method is robust.

The comparison against baselines raises methodological questions. The authors compare Flow-VQE against GD, Adam, and QNSPSA with fixed hyperparameters, but hyperparameter optimization is standard practice for baseline comparisons. The statement “All baseline optimizers use a learning rate of =0.02” in Figure 2 suggests baselines were not individually tuned, potentially handicapping them. Additionally, the QNSPSA comparison is

welcome (as a gradient-free method), but the manuscript does not include other recent gradient-free approaches such as COBYLA, Nelder-Mead, or evolutionary strategies that are commonly used in VQE literature.

Reproducibility infrastructure is not adequately described. The manuscript mentions “state-vector simulation experiments” using PennyLane and OpenFermion but does not provide: (a) version numbers for software dependencies, (b) hardware specifications for classical computation, (c) wall-clock training times, or (d) code/data availability statements beyond acknowledging NAISS computational resources. For a method whose practical utility depends on the classical overhead remaining “easily tractable with modern machine learning techniques,” timing data would strengthen the claims. The required npj Quantum Information statements (Code Availability, Data Availability) appear to be missing from the current draft.

Questions for Authors: 1. Can you report multi-seed variance (n_5) for the key comparisons in Figures 2-4 and Table I? 2. What ablations can you provide for buffer size M , batch size B , flow depth, and regularization noise? 3. Will code and data be released upon publication, and at what URL? 4. What are the wall-clock training times for Flow-VQE-S and Flow-VQE-M on a specified hardware configuration?

Verdict: 5/10 — Major Revisions

5.2.4 Voice 4 — Devil’s Advocate

This manuscript exemplifies a troubling trend in quantum computing: impressive-sounding improvements over carefully chosen baselines that dissolve under scrutiny. Let me enumerate the fundamental problems that my fellow reviewers have been too generous about.

The baseline comparisons are rigged. The entire narrative hinges on comparing against gradient descent, an optimizer no serious VQE practitioner would use for production work. The authors bury the admission that “Flow-VQE-S achieves approximately a two- to five-fold reduction” over Adam—this is the *only* meaningful comparison, and 2-5 \times improvement is incremental, not transformative. The “two orders of magnitude” headline claim requires comparing against raw GD at $\eta=0.02$ without momentum, which is a strawman. Furthermore, baseline optimizers were run with identical learning rates rather than individually tuned, while Flow-VQE’s hyperparameters (learning rate, weight decay, flow depth, buffer size, batch size, noise variance) were presumably chosen via undocumented empirical search. This asymmetry invalidates the entire quantitative comparison.

The method cannot scale. The authors test on systems with 5-12 qubits and 54-117 parameters. Modern quantum chemistry requires hundreds to thousands of qubits. Section VI’s optimistic claim that “the classical overhead in modern normalizing-flow architectures scales linearly in d ” ignores the elephant in the room: flow models require $O(d)$ samples to estimate the target distribution in d dimensions, and preference-based training with buffer size $M=2$ cannot possibly capture a meaningful distribution over 10,000+ parameters. The authors’ own Figure 6 shows extensive stagnation plateaus even at $d=100$ —these will become impassable walls at chemically relevant scales. The comparison against parameter transfer (Figure 4) shows Flow-VQE-M only marginally outperforming PT, which requires no neural network overhead whatsoever.

The experimental design hides failures. Why are only four molecules tested? Why these specific geometric distortions? The authors test HO bond stretching, H linear chain stretching, NH inversion, and CH C-H stretch—but conspicuously absent are: (a) molecules with transition metals, (b) open-shell systems, (c) excited states, (d) strongly correlated systems beyond stretched bonds, and (e) any molecule requiring >12 qubits. The selection bias suggests the authors tried other systems and failed. The manuscript contains no honest negatives, no failure modes section, and no discussion of when Flow-VQE underperforms baselines. Figure 2(b) shows Flow-VQE-S *losing* to Adam at H 0.6 Å—this is mentioned in passing but not analyzed. A rigorous empirical study would characterize the conditions under which the method fails.

The novelty is overstated. The core idea—use generative models to propose VQE parameters—appears in multiple prior works: Ceroni et al. (Ref. [34], 2023), Zhang et al. (Ref. [35], 2025), and Nakaji et al. (Ref. [80], 2024). The preference-based training draws directly from DPO in language models without significant adaptation for quantum settings. The Gaussianization flow architecture is off-the-shelf (Meng et al., AISTATS 2020). The conditional embedding via Hamiltonian coefficients is straightforward. Stripping away the quantum framing, this is a routine application of conditional generative modeling to a black-box optimization problem—technically competent but not meeting the novelty bar for a flagship publication.

The required journal elements are missing. npj Quantum Information mandates Author Contributions, Competing Interests, Data Availability, and Code Availability statements. None appear in this draft. The

Methods section is scattered throughout rather than consolidated at the end per Nature/npj conventions. The abstract (currently 248 words) barely meets the 250-word limit. These are not fatal flaws but indicate either carelessness or unfamiliarity with the target venue.

Recommendation: Reject. The contributions do not justify publication in npj Quantum Information. The method shows modest improvements over properly chosen baselines, cannot scale to relevant system sizes, lacks statistical rigor, and does not advance the field beyond concurrent/prior work. The manuscript may be suitable for a computational chemistry venue after substantial revision.

5.2.5 Voice 5 — Editor-in-Chief synthesis

Having carefully considered all four reviews, I observe substantial disagreement regarding the manuscript’s merits. Reviewer 1 finds the physics sound with minor concerns about numerical precision and ansatz expressibility. Reviewer 2 identifies legitimate novelty in the preference-based training scheme but requests stronger baseline comparisons and engagement with concurrent work. Reviewer 3 raises serious methodological concerns about statistical rigor and reproducibility that must be addressed. The Devil’s Advocate presents the strongest case for rejection, emphasizing baseline selection bias, scalability limitations, and missing journal requirements.

Let me address the Devil’s Advocate’s critiques specifically, as they represent the highest bar the authors must clear. The baseline comparison criticism has merit: comparing against GD is indeed uninformative, and the Adam comparison should be the primary benchmark. However, the 2-5E improvement over Adam, combined with the generative warm-start capability (Table I), does represent meaningful practical utility—especially for practitioners who must optimize many related molecular configurations. The scalability criticism is fair but applies to essentially all NISQ algorithm papers; the authors appropriately scope their claims to near-term devices. The charge of hidden failures is partially addressed by the H 0.6 Å result where Flow-VQE underperforms, though I agree a dedicated failure analysis would strengthen the work. The novelty critique is the most serious: the manuscript must better differentiate from Nakaji et al. (GQE) and demonstrate why parameter-space generation offers advantages over circuit-space generation.

The statistical concerns raised by Reviewer 3 are non-negotiable for publication. Single-seed results without confidence intervals are insufficient for empirical claims. The missing Data/Code Availability statements violate npj QI policy and must be added. The incomplete ablation study leaves readers unable to assess which design choices are essential versus incidental.

Regarding the physics and chemistry (Reviewer 1), the concerns are addressable through clarifications rather than new experiments. The algorithmic novelty (Reviewer 2) requires engagement with concurrent literature but does not fundamentally undermine the contribution. The preference-based training adapted from RLHF to quantum chemistry contexts is a genuine methodological contribution, even if the components are individually known.

Final Verdict: Major Revisions

The manuscript presents a technically sound method with meaningful practical utility for VQE optimization across molecular configurations. However, the current presentation suffers from inflated claims relative to appropriate baselines, inadequate statistical rigor, missing journal-required elements, and insufficient differentiation from concurrent work. Acceptance is possible after substantial revision.

Must-fix items before resubmission (ordered by severity):

1. **Add multi-seed variance reporting** (n5) for all quantitative comparisons, including Figures 2-4 and Table I. Report 95% confidence intervals or standard deviations.
2. **Include Code and Data Availability statements** per npj QI requirements, with repository URLs and DOIs where applicable.
3. **Revise headline claims** to foreground Adam comparisons (2-5E improvement) rather than GD comparisons (100E). Move GD results to supplementary material or present as secondary.
4. **Add direct comparison with GQE** (Nakaji et al.) or explicitly justify why parameter-space generation is preferable to circuit-space generation for the tested molecules.
5. **Provide ablation study** for buffer size M, flow depth, and regularization noise, minimally as supplementary material.

6. **Consolidate Methods section** at the end of the manuscript per Nature/npj conventions, and add required Author Contributions and Competing Interests statements.
7. **Specify numerical precision** (complex64/128) and verify that precision artifacts do not affect comparisons at the 1.6 mHa threshold.
8. **Add failure mode analysis** discussing conditions under which Flow-VQE underperforms baselines, expanding on the H 0.6 Å observation.

5.2.6 Vote table

Voice	Recommendation	Confidence 1-10
Reviewer 1	Minor Revisions	7
Reviewer 2	Major Revisions	6
Reviewer 3	Major Revisions	8
Devil's Advocate	Reject	7
Editor-in-Chief	Major Revisions	8

5.3 flowvqe — Logical-fallacy report

Model: claude-haiku-4-5-20251001 · 2,863 output tokens · 52.8 s

I'll carefully analyze this quantum computing manuscript for logical fallacies using the provided taxonomy, focusing on findings with severity at or above **medium**.

Fallacy: cherry-picked-baseline **Severity:** high **Location:** Section I (Introduction) and Section V.A (Single-molecule optimization), Figure 2 **Evidence:** “We compare Flow-VQE to a number of standard benchmarks through numerical simulations. . . We find that Flow-VQE outperforms baseline optimization algorithms” and comparing against “gradient descent (GD), quantum natural-gradient simultaneous perturbation stochastic approximation (QNPSA), Adam” **Why it’s the fallacy:** The paper compares against generic optimization algorithms (GD, Adam, QNPSA) but omits comparison with other published warm-start and ML-based VQE initialization methods that they explicitly cite as prior work (references [32-44]), including supervised learning approaches, meta-learning frameworks, and other generative modeling methods. For instance, they cite methods like FLIP [42] and meta-VQE [41] but don’t benchmark against them on the same molecular systems. **Suggested fix:** Include direct numerical comparisons with at least one or two of the most relevant published warm-start methods (e.g., the supervised learning approaches from [32-33] or the meta-learning methods from [40-44]) on the same Hamiltonians, or explicitly state why such comparisons were not feasible and acknowledge this as a limitation.

Fallacy: conflated-regimes **Severity:** high **Location:** Section I (Introduction), Section VI (Limitations and Future Work) **Evidence:** “While the numerical experiments reported here extend only to 12-qubit active spaces and 117 variational parameters, several features of Flow-VQE promise broad quantum-resource savings, even when scaling to larger molecules.” **Why it’s the fallacy:** The paper extrapolates from small Hamiltonians (5-12 qubits, up to 117 parameters) to claims about larger systems without empirical validation. The claim that advantages will persist at scale is speculative, especially given that barren plateaus and optimization landscape complexity scale non-trivially with system size. The paper acknowledges this only partially in limitations but still makes forward-looking claims about scalability. **Suggested fix:** Temper scalability claims with explicit caveats. Replace “promise broad quantum-resource savings” with “may potentially offer resource savings pending empirical validation at larger scales” and acknowledge that the scaling behavior of the normalizing flow approach with qubit count remains uncharacterized.

Fallacy: hasty-generalization **Severity:** medium **Location:** Section V.C (Estimate of cost advantage) **Evidence:** “For NH₃, standard VQE requires an average of CVQE = 5,265 circuit evaluations per test point. . . These estimates are instance-dependent and not intended as universal benchmarks, but they illustrate the practical advantages of Flow-VQE-M” **Why it’s the fallacy:** The cost advantage analysis is based on only two molecules (NH₃ and C₆H₆) with very limited training configurations (4 configurations each). The paper then uses these narrow results to draw broader conclusions about the method’s advantages, despite acknowledging the estimates are “instance-dependent.” **Suggested fix:** Explicitly state that the cost advantage calculations are illustrative examples from a narrow test set, not generalizable predictions. Add language such as: “These results demonstrate potential advantages in specific cases but should not be extrapolated to other molecular systems without further validation.”

Fallacy: active-space-handwave **Severity:** medium **Location:** Section IV.A.1 (Electronic structure modeling) and Section VII (Conclusion) **Evidence:** “For active-space selections, we perform them to manage computational complexity while preserving essential electronic structure features: (4e, 4o) for H₄, (6e, 5o) for H₂O, (6e, 6o) for NH₃, and (6e, 6o) for C₆H₆” combined with conclusion claims that “Flow-VQE can become a pragmatic and versatile paradigm” **Why it’s the fallacy:** The paper uses small, carefully selected active spaces but claims generalization without running on larger or different active space selections. The choice of active spaces is not justified beyond “managing computational complexity,” and there’s no evidence the method would work for larger active spaces that would be needed for chemically meaningful calculations on these molecules. **Suggested fix:** Add explicit justification for why these specific active spaces were chosen and acknowledge that performance on larger, more chemically realistic active spaces remains untested. Include a statement like: “The active spaces used here are minimal and primarily serve as proof-of-concept; extension to larger active spaces required for chemical accuracy remains to be demonstrated.”

Fallacy: hardware-irrelevant-comparison **Severity:** medium **Location:** Section IV (Numerical Simulations), throughout **Evidence:** “We empirically validate Flow-VQE through state-vector simulation experiments on various quantum chemical systems” and “We adopt the number of quantum circuit evaluations, independent of measurement shot counts, as the primary performance metric” **Why it’s the fallacy:** All experiments are performed on ideal state-vector simulators without noise calibration or hardware validation. The paper claims the method is designed for “near-term quantum devices” and NISQ hardware, but provides no evidence of performance under realistic noise conditions. Circuit evaluation counts on simulators don’t translate directly to hardware performance where noise accumulates. **Suggested fix:** Either include noisy simulations with realistic noise models, or clearly caveat that all results are for ideal simulators and that hardware validation is required before claims about NISQ utility can be made. Add explicit text: “All experiments were conducted under ideal noiseless conditions; performance under realistic hardware noise remains to be characterized.”

Fallacy: asymptotic-only-claim **Severity:** medium **Location:** Section VI (Limitations and Future Work) **Evidence:** “On the classical side, the classical overhead in modern normalizing-flow architectures scales linearly in d , keeping training and inference practical even as d reaches the tens of thousands” **Why it’s the fallacy:** The paper claims linear scaling to “tens of thousands” of parameters but only demonstrates results up to $d=117$ parameters. This asymptotic claim about classical overhead is not validated empirically in the presented work. **Suggested fix:** Either remove the claim about scaling to tens of thousands of parameters, or explicitly note this is a theoretical expectation that has not been validated: “While normalizing flow architectures theoretically scale linearly in d , we have only validated this up to $d=117$ in the present work.”

Fallacy: unit-inflation **Severity:** medium **Location:** Section I (Abstract), Section V (Results) **Evidence:** “improvements range from modest to more than two orders of magnitude” and “up to 50-fold compared with

Hartree–Fock initialization” **Why it’s the fallacy:** The paper presents maximum improvements (“up to,” “more than two orders of magnitude”) rather than typical or median improvements. The most dramatic improvements (50-fold, 100x) occur at specific learning rates (=0.001) or extreme bond lengths that may not represent typical use cases. The “two orders of magnitude” claim appears to come from outlier configurations. **Suggested fix:** Report median or geometric mean improvements alongside maximum improvements. Replace “up to 50-fold” with “improvements ranging from X-fold to 50-fold (median: Y-fold)” and clarify that the most dramatic improvements occur under specific conditions.

5.4 flowvqe — Stage-6 CQE narrative

Model: claude-opus-4-5-20251101 · 1,180 output tokens · 31.7 s

Process Summary: QuantumNovelty Run Evaluation

5.4.1 Composite Verdict

The composite score of **23/100** places this run firmly in the “Critical Deficiencies” tier. On the standard 1-100 scale, scores below 30 indicate a run that failed to produce the fundamental artifacts required for a credible scientific claim. This is not a borderline result—it represents a systematic failure across nearly all evaluation dimensions.

A geometric mean of 23 from six dimensions means no single strong dimension could compensate for the others. The multiplicative nature of the composite punishes runs where any dimension collapses toward zero, and here we see scores of 8, 20, 27, 30, 30, and 40. The lowest scores drag the composite down irreversibly. This run produced neither the baseline comparisons nor the verification artifacts that would allow anyone—including the original researchers—to assess whether the results mean anything.

5.4.2 Strongest Dimension: Communication (40)

Communication scored highest at 40, though “highest” is relative when the ceiling was 40 and the floor was 8. Both probes—**logical fallacies absent** and **reviewer panel verdict**—scored 40, but notably with evidence showing neither check was actually run. The `logical_fallacies` skill was never invoked, and no `review_panel.md` was generated.

This score reflects an absence of detected problems rather than a presence of verified quality. The run didn’t produce obviously fallacious reasoning, but it also didn’t subject itself to the adversarial scrutiny that would surface subtle issues. A score of 40 here essentially means “we didn’t catch you making errors because we didn’t look.” This is the evaluation equivalent of passing a test you never took.

What this reveals about the run: the team may have prioritized moving forward over pausing to validate. Communication artifacts are often treated as polish to be added later, but without a review panel verdict or fallacy check, there’s no evidence the core claims were stress-tested before being considered complete.

5.4.3 Weakest Dimension: Novelty Rigour (8)

At 8/100, Novelty Rigour represents the critical failure mode of this run. The two probes tell a damning story:

- **Augmented baseline catalog present** scored 10, with evidence showing “`baseline_catalog` has 0 rows.” This means the run produced zero baseline comparisons. Without a catalog of known results, there is no way to determine whether any finding represents genuine novelty or rediscovery.
- **Strict-domination comparator run** scored 5, with “`novelty_verdict.json` not found.” The comparator that would determine whether results strictly dominate known baselines was never executed.

The stage that produced this failure was clearly the **baseline establishment phase**. Before any novelty claims can be made, the run must populate a baseline catalog with prior art—known molecular energies, established operator counts, published gate complexities. This catalog wasn’t just incomplete; it was empty.

Zero rows means zero comparisons were possible. The strict-domination comparator couldn't run because there was nothing to compare against.

This is a foundational failure. Everything downstream—claims of improvement, assertions of novelty, excitement about results—rests on sand. The run may have produced interesting outputs, but without baseline comparisons, “interesting” is indistinguishable from “already known” or even “worse than prior work.”

5.4.4 Three Highest-Leverage Improvements

1. Populate the Baseline Catalog Before Starting Discovery The single highest-leverage fix is requiring a non-empty baseline catalog as a gate before the novelty search begins. This means compiling known results from literature, prior runs, or established benchmarks for the target molecules or problems. The baseline catalog should include at minimum: molecule identifiers, active space configurations, best-known energies with citations, and operator/gate counts from published methods. A run with 50 baseline rows would have transformed this evaluation.

2. Generate the Structured Pareto Archive The **pareto archive structured** probe scored 20 with evidence showing “archive rows: 0.” A Pareto archive captures the trade-off frontier (e.g., accuracy vs. circuit depth) and preserves non-dominated solutions. Without this, there's no structured record of what the run explored. Implementing automatic Pareto archiving during the search loop—appending each non-dominated solution with its objective values—would improve Reproducibility and provide the raw material for novelty assessment.

3. Run the Strict-Domination Comparator and Generate `novelty_verdict.json` Even with a sparse baseline catalog, running the comparator produces a verdict file that honestly reports “no novelty proven” or “insufficient baselines.” This file forces the run to confront its evidential status. The improvement here is procedural: make `novelty_verdict.json` generation a mandatory step before any run is considered complete. A verdict of “cannot assess” is more valuable than an absent verdict, because it surfaces the gap explicitly.

This run demonstrates what happens when execution outpaces verification infrastructure. The machinery to find novel solutions may have operated, but the machinery to prove they're novel was never built. A score of 23 is recoverable—but only if the next run treats baseline establishment and novelty verification as prerequisites, not afterthoughts.

6 hwqml — *Trainability and Expressivity of Hamming-Weight Preserving Quantum Circuits for Machine Learning*

Léo Monbroussou, Elliott Z. Mamon, Jonas Landman, Alex B. Grilo, Romain Kukla, Elham Kashefi [arXiv:2309.15547](https://arxiv.org/abs/2309.15547)
Quantum 9, 1745 (2025)

6.1 hwqml — Deep-research review

Model: claude-opus-4-5-20251101 · 1,546 output tokens · 42.7 s

6.1.1 1. One-paragraph summary of what the paper claims

This paper analyzes the trainability and controllability of Hamming-weight (HW) preserving variational quantum circuits (VQCs) for quantum machine learning. The authors make three main contributions: (1) they design and prove the feasibility of new heuristic data loaders that perform quantum amplitude encoding of (n choose k)-dimensional vectors using n -qubit circuits made of Reconfigurable Beam Splitter (RBS) or Fermionic Beam Splitter (FBS) gates, with existence proofs based on controllability arguments via the Quantum Fisher Information Matrix (QFIM) rank; (2) they prove that the rank of the QFIM for any VQC state is almost-everywhere constant (Theorem 1); and (3) they provide trainability analysis showing that the variance of the l_2 cost function gradient scales as $O(1/(n \text{ choose } k))$, proving conditions for existence/absence of Barren Plateaus for these circuits. Notably, they claim this represents a counterexample to a conjecture from [11] linking controllability (via DLA dimension) to trainability, since HW-preserving circuits can avoid Barren Plateaus even without full controllability or 2-design assumptions.

6.1.2 2. Audit-and-falsify checklist

- Augmented baseline catalog:** PARTIAL — The paper compares against the theoretical framework of [11] (Larocca et al.) and discusses relationships to [17, 18, 25], which are contemporary works on DLA-trainability connections. However, for the data loading/encoding component, comparisons are primarily to [5] (the prior HW-preserving loader in unary basis) without systematic benchmarking against other amplitude encoding methods.
- Strict-domination comparator:** NOT-APPLICABLE — The paper does not make Pareto optimality claims; the main results are theoretical (existence proofs, variance bounds) rather than empirical performance comparisons requiring tolerance calibration.
- Recompute-from-raw:** PARTIAL — Figures 4, 6, 9, and 10 display numerical results, and Table 2 shows simulation errors. The theoretical predictions (dotted lines in Figs. 9-10) appear consistent with the plotted data, but no raw data files or recomputation scripts are referenced. The paper text indicates the simulation was done via Numpy/Qiskit but no audit trail is provided.
- Wilson 95% CIs:** FAIL — Table 2 reports “Average Error” and “Variance” for 1000 samples but does not provide confidence intervals. Figures 9-10 show variance comparisons but without error bars or uncertainty quantification on the numerical estimates.
- Cross-LLM falsifiability:** NOT-APPLICABLE — No LLM-in-the-loop methodology is employed; this is a theoretical/numerical quantum computing paper.
- Honest negatives:** PARTIAL — The paper acknowledges limitations: FBS gates have reduced controllability (Section 2.2, Appendix C), classical simulability limits speedup (Section 3.3), and Theorem 2 relies on an unproven conjecture (Conjecture F1, acknowledged in Section 3.2). However, there is no dedicated “Failure Modes” section, and no cases where the proposed data loader fails to converge are reported.
- Simulator precision floor:** NOT-APPLICABLE — The paper does not compare quantum vs. classical energy calculations; numerical simulations are for gradient variance and QFIM rank, which are computed symbolically/analytically where possible.

- **Auditable claims:** FAIL — No `audit_claims.py` or equivalent is provided. The paper does not reference a code repository, reproducibility package, or JSON files for numerical claims. Quantum journal requires Code Availability statements, but this appears as an arXiv preprint accepted to Quantum without such artifacts being explicitly linked.

6.1.3 3. Overall assessment

This paper presents rigorous theoretical work on an important topic (trainability of symmetric VQCs), with novel analytical results that bypass the standard 2-design assumptions used in prior work. The mathematical development is careful, with proofs relegated to extensive appendices. The main weakness from an audit-and-falsify perspective is **reproducibility infrastructure**: the numerical claims in Figures 4, 6, 9-10 and Table 2 lack uncertainty quantification, and no code/data artifacts are referenced for independent verification. The reliance on Conjecture F1 for Theorem 2 is honestly disclosed but represents an unresolved gap. The theoretical claims themselves appear internally consistent, and the counterexample to Conjecture 1 from [11] is well-argued. However, strict reviewer-mode auditing would flag the absence of confidence intervals on small-sample statistics and the lack of reproducibility artifacts as significant concerns for a journal requiring Code/Data Availability statements.

Research rigour score: 6/10 — Strong theoretical development with honest acknowledgment of limitations, but lacking the reproducibility infrastructure and statistical rigor expected under the audit-and-falsify framework.

6.1.4 4. Three highest-leverage improvements

1. **Add uncertainty quantification to all numerical results:** For Figures 9-10, report bootstrap or Wilson 95% confidence intervals on the estimated variance values. For Table 2, add confidence intervals on the mean error. The claim “variance follows theoretical values” in Figures 9-10 should be supported by statistical tests (e.g., χ^2 goodness-of-fit against the predicted $8k(nk)/(n(n-1)d_k)$ values).
2. **Provide a reproducibility package with `audit_claims.py`:** Create a repository containing (a) code to regenerate all figures from raw simulation data, (b) JSON files storing the numerical data points underlying Figures 4, 6, 9-10 and Table 2, and (c) a script that re-derives every numerical claim (e.g., QFIM ranks, gradient variances, classification accuracies) from these files. This satisfies Quantum journal’s Code Availability requirement and enables independent verification.
3. **Either prove Conjecture F1 or provide explicit numerical bounds:** Theorem 2 currently depends on Conjecture F1 about spectral gaps of stochastic matrices. Either (a) complete the proof (the numerical evidence in Appendix F.9 is suggestive but not definitive), or (b) provide explicit numerical values for the spectral gaps of T matrices arising from specific connectivities (e.g., nearest-neighbor, Rigetti ASPEN M2) with rigorous error bounds, so that the theorem’s applicability can be verified for concrete cases.

6.2 hwqml — Reviewer panel 5 voices

Model: claude-haiku-4-5-20251001 · 5,296 output tokens · 124.3s

6.2.1 Voice 1 — Reviewer 1 (Physics correctness)

This paper investigates Hamming-weight preserving variational quantum circuits, specifically those constructed from Reconfigurable Beam Splitter and Fermionic Beam Splitter gates, analyzing their controllability and trainability properties. The physics foundation of the work rests on the observation that these gates preserve subspaces of fixed Hamming weight, reducing the effective Hilbert space dimension from 2^n to the binomial coefficient $\binom{n}{k}$. The authors correctly identify that the RBS gate implements a planar rotation between computational basis states $|01\rangle$ and $|10\rangle$ as shown in Equation 3, which is indeed the standard form of this gate used in photonic and fermionic quantum computing platforms. The Hamiltonian HRBS in Equation 4 is

correctly constructed as the generator of this rotation, though I note the authors do not explicitly verify that $\exp(-iHRBS)$ produces the stated unitary form, which would strengthen the presentation.

The treatment of the Fermionic Beam Splitter in Definition 4 deserves particular scrutiny. The authors introduce the fermionic parity factor $f = \prod_{i < j} s_{ij}$, which accounts for the anticommutation relations of fermionic operators when mapped to qubits via Jordan-Wigner transformation. This is physically correct and represents the key distinction between bosonic and fermionic systems. However, the paper’s claim in Appendix C that “each block W^k for $k > 1$ is perfectly determined by W^1 as it is the k -compound matrix of W^1 ” requires more careful justification. While this relationship holds for the fermionic case due to the determinantal structure of fermionic wavefunctions, the authors should clarify whether this restricts the FBS architecture to Jordan-Wigner ordering specifically, or whether the result generalizes to other fermion-to-qubit mappings such as Bravyi-Kitaev or parity encoding. The DLA dimension upper bound of $n(n-1)/2$ for FBS circuits stated in Figure 4 appears consistent with this compound matrix structure, but the connection to physical symmetries (particle number conservation) could be made more explicit.

The Quantum Fisher Information Matrix analysis in Section 2.3 is mathematically sound, with the definition in Equation 9 matching standard formulations. The authors’ claim in Theorem 1 that the QFIM rank is almost-everywhere constant follows from standard results on analytic functions and Jacobian ranks, though the proof in Appendix D could benefit from more explicit connection to the VQC structure. One technical concern is the numerical precision of QFIM rank calculations referenced in Figure 6: the authors should specify whether they used complex128 or complex64 arithmetic, as rank determination near singular matrices is notoriously sensitive to floating-point precision. The claim that “drawing a point uniformly at random and calculating its QFIM rank yields $r() = r_{\max}$ with probability 1” is mathematically correct but practically requires discussion of numerical tolerances used in rank determination.

The trainability analysis in Section 3 connects the gradient variance to the subspace dimension through Lemma 1, with the key result that gradients scale as $O(1/d_k)$ rather than exponentially. The backpropagation formulation in Equations 15 and 16 correctly accounts for the multi-pair rotations induced by single gates in higher Hamming-weight subspaces. However, I am concerned about the cost function choice in Equation 13, which uses squared Euclidean distance between quantum states. The authors acknowledge in a footnote that this is not directly expressible as an observable expectation value, but the proposed ancilla-based workaround from reference [6] adds circuit depth that could affect the practical trainability claims. The proofs in Appendices E through G are technically detailed and appear correct, though the reliance on Conjecture F1 regarding spectral gaps of stochastic matrices represents an acknowledged gap in the rigorous chain of reasoning.

Questions for Authors: 1. Does the compound matrix relationship for FBS gates hold for Bravyi-Kitaev mapping, or is Jordan-Wigner ordering implicit throughout? 2. What numerical tolerance was used for QFIM rank determination in Figure 6, and how sensitive are the reported ranks to this choice? 3. Can you provide any analytical progress on Conjecture F1, or at minimum characterize the class of patterns for which it is provably true?

Verdict: 7/10. Recommendation: minor-revisions.

6.2.2 Voice 2 — Reviewer 2 (Algorithmic novelty)

The central contributions of this work must be evaluated against the rapidly evolving landscape of trainability analysis for variational quantum circuits. The authors position their work relative to Larocca et al. (2022, reference [11]), Ragone et al. (2024, reference [17]), and Fontana et al. (2024, reference [18]), claiming independence from these approaches and applicability in settings where the DLA-based framework does not apply. This positioning is partially justified: the recent work by Ragone et al. established that variance of cost gradients scales inversely with the DLA dimension under 2-design assumptions, while Fontana et al. extended this to include the adjoint representation framework. Both papers explicitly exclude or provide only upper bounds for Hamming-weight preserving circuits, and the present work fills this gap by providing exact variance formulas for RBS/FBS circuits without invoking 2-design assumptions.

However, the novelty must be tempered by examination of concurrent and recent work. Diaz et al. (2023, reference [25]) specifically addressed scenarios evading the DLA framework, providing theory that the authors acknowledge as “consistent” with their results. More critically, the data loading application in Section 2 closely parallels the approach of Johri et al. (2021, reference [5]) and Landman et al. (2022, reference [6]), which already demonstrated HW-preserving amplitude encoding in the unary basis. The extension to arbitrary Hamming

weight k represents an incremental generalization rather than a conceptual breakthrough. The controllability analysis via QFIM is also not novel—Haug et al. (2021, reference [29]) established the connection between QFIM rank and circuit capacity, and the algorithmic approaches in Section 2.3 (Algorithms 1 and 2) are straightforward applications of greedy rank-maximization strategies that have appeared in quantum control literature for decades.

The trainability results themselves, while technically sound, raise questions about practical relevance. The main theoretical contribution is that gradient variance scales as $8k(n-k)/(n(n-1)d_k)$ per Theorem 3, avoiding exponential barren plateaus for fixed k . However, this polynomial scaling with $d_k = \binom{n}{k}$ still implies vanishing gradients when k is not held constant. The authors acknowledge this in Section 3.3 but do not adequately address the practical regime of interest. For quantum machine learning applications requiring expressive circuits, small k implies limited expressivity (the authors themselves note in Section 2.2 that reduced k may be necessary “in hopes of achieving full controllability for a smaller subspace”). This creates a fundamental tension between trainability and expressivity that the paper does not resolve, and which was already identified in Holmes et al. (2022, reference [30]). The claimed absence of barren plateaus is therefore conditioned on operating in a regime that may be of limited practical interest.

Regarding computational comparisons, the classical simulability of HW-preserving circuits in polynomial-dimension subspaces is well-established (the authors cite Anschuetz et al. 2023, reference [27]). The running time analysis in Table 1 is straightforward but the claimed “polynomial advantage” for quantum training is misleading: if the circuit is classically simulable, quantum training offers no asymptotic speedup, only constant-factor improvements from parallelization of gate applications. The suggestion that one could “train them classically to represent classical data and then associate them with a quantum circuit that is harder to simulate” (end of Section 2.3) is speculative and would require a separate analysis of how trainability transfers across the transition from simulable to non-simulable regimes. This gap between the theoretical guarantees and practical quantum advantage undermines the paper’s positioning as relevant to near-term quantum machine learning.

The comparison with the Conjecture 1 from Larocca et al. is interesting but overstated. The authors claim to “refute” this conjecture “in its full generality,” but more accurately, they demonstrate a setting where the conjecture’s hypotheses are not satisfied (HW-preserving circuits do not satisfy the full controllability assumption), so the conjecture simply does not apply. This is not a refutation but rather a delineation of the conjecture’s domain. The positive contribution is showing that trainability can be analyzed without invoking the conjecture’s framework, but this does not invalidate the conjecture for circuits where it does apply.

Questions for Authors: 1. For what range of k values do you expect the polynomial gradient scaling to remain practically useful for gradient-based optimization? 2. Can you quantify the “polynomial advantage” claimed for quantum training in concrete circuit parameters? 3. How do your trainability guarantees extend when HW-preserving blocks are combined with non-HW-preserving operations, as would be needed for practical applications?

Verdict: 6/10. **Recommendation:** major-revisions.

6.2.3 Voice 3 — Reviewer 3 (Empirical evidence)

The empirical validation presented in this paper is concentrated in Section 4 and the accompanying figures, with the bulk of the technical content being theoretical. While the mathematical results are substantial, the numerical evidence supporting these results suffers from several methodological shortcomings that limit confidence in the practical applicability of the findings. The authors present simulations but do not provide the statistical rigor expected in modern quantum computing research, particularly given the stochastic nature of VQC training and the sensitivity of gradient-based claims to initialization and sampling.

Figure 6 presents the evolution of QFIM rank for a periodic structure ansatz, but the methodology for rank determination is not specified. The authors claim to verify Theorem 1 (“rank is almost-everywhere constant”) numerically, but the figure shows only single point evaluations without error bars or repeated sampling. Given that numerical rank computation is notoriously sensitive to condition number and floating-point precision, the authors should report the singular value spectrum and the threshold used to distinguish zero from nonzero singular values. The statement “upon randomly sampling parameter values $[0, 2]^p$, the value of $\text{rank}[\text{QFIM}()]$ is independent of ” requires empirical verification over multiple samples with statistical characterization of any observed variation. The claim that one random sample suffices to determine maximum rank (Lemma D3) is mathematically correct but practically requires confidence intervals on the numerical rank determination.

The gradient variance simulations in Figures 9 and 10 are more carefully executed, with 100 and 10000 random samples respectively. However, the agreement with theoretical predictions (dotted lines) is claimed but not quantified. The authors should report chi-squared goodness-of-fit statistics or at minimum the maximum observed deviation from theory. The figures show reasonable agreement but there appear to be systematic deviations at higher L values in Figure 9 panels (3) and (4) that are not discussed. More concerning, the experimental setup for Figure 10 fixes the initial state as a basis state rather than sampling uniformly, which the authors acknowledge places this “in between the assumptions of spherical t -designs of $t = 1$ and $t = 2$.” This makes direct comparison with Theorem 3’s predictions problematic, and the claim that “points only roughly follow the dashed $1/d_k$ values” is a red flag for potential systematic effects that are not captured by the theory.

The Fashion MNIST simulations in Section 4.1 and 4.4 represent the only realistic application test in the paper, but the experimental design is inadequate for drawing meaningful conclusions. Table 2 reports mean error and variance over 1000 samples for a single circuit configuration, without comparison to baseline methods or random initialization baselines. The binary classification results in Figure 11 show learning curves but do not report final accuracy values, number of training epochs, or comparison to classical neural networks of equivalent parameter count. The claim that “we do not claim that this plot exhibits any advantage to use RBS based quantum circuits as neural networks” is appropriate but undermines the relevance of including this experiment at all. If the goal is merely to demonstrate that training converges, this should be stated explicitly with appropriate caveats about the toy nature of the demonstration.

The reliance on Conjecture F1 for the main theoretical results (Theorem 2) is acknowledged but inadequately supported. The numerical evidence in Appendix F.9 examines only three specific patterns (line-down, line-downup, pyramid) for $n = 50$ qubits and $k \in \{1, 2, 3\}$. The polynomial decay fits show r^2 values above 0.999, which is encouraging, but the fitted exponents (slopes of -1.99, -1.96, -0.95 in Figure F1) vary significantly across patterns, suggesting non-universal behavior that the conjecture does not capture. The authors do not examine patterns with higher connectivity or larger gate counts, which would be more representative of practical circuit designs. Furthermore, the conjecture requires spectral gap bounds that scale as $(1/\text{poly}(n))$, but the specific polynomial is not determined from the numerical fits, leaving the dependence on circuit architecture as an open question that directly affects the required circuit depth for Theorem F4 to apply.

Questions for Authors: 1. What singular value threshold was used for QFIM rank determination, and how does changing this threshold affect the reported ranks? 2. Can you provide statistical tests (chi-squared, Kolmogorov-Smirnov) for the agreement between observed gradient variances and theoretical predictions in Figures 9-10? 3. For Conjecture F1, can you characterize the polynomial exponent as a function of pattern properties, or at minimum provide bounds on the constants involved? 4. Is the code for reproducing all numerical results publicly available, and if so, does it include scripts for regenerating all figures from raw data?

Verdict: 5/10. Recommendation: major-revisions.

6.2.4 Voice 4 — Devil’s Advocate

This paper represents a competent but ultimately incremental contribution that has been dressed up with elaborate mathematical machinery to obscure fundamental limitations. Let me be direct about the critical weaknesses that my colleagues have treated too gently.

The core claim of avoiding barren plateaus is technically true but practically vacuous. The trainability guarantee requires keeping the Hamming weight k fixed as the number of qubits n grows, which means the subspace dimension $d_k = \binom{n}{k}$ scales polynomially with n . But this polynomial scaling still implies gradient variance decreasing as $8k(n-k)/((n-1)d_k)$, which for $k=2$ becomes $O(1/n^2)$. Even avoiding exponential vanishing, a cubic decrease in gradient magnitude renders optimization impractical for $n > 50$ qubits without shot counts scaling as n to maintain fixed signal-to-noise. The authors conflate “not exponentially vanishing” with “trainable,” which is a fundamental category error. The theoretical contribution is that gradients vanish polynomially rather than exponentially, but this distinction matters only if the polynomial degree is small enough for practical optimization, which the authors never establish.

The dependence on Conjecture F1 is more problematic than acknowledged. The conjecture asserts that spectral gaps of stochastic matrices associated with connected RBS/FBS patterns scale as $(1/\text{poly}(n))$, but the numerical evidence in Appendix F.9 examines only three elementary patterns for modest qubit counts. The

fitted decay exponents vary from -0.95 to -1.99 across patterns, suggesting the polynomial degree is architecture-dependent in ways not captured by the conjecture. More critically, the authors do not bound the constants in this polynomial scaling, so even if the conjecture holds, the required circuit depth $L \propto d_k^n$ from Theorem F4 could be astronomically large for circuits of practical interest. The condition that angles must be located “at any constant fraction of the depth” means trainability is guaranteed only for gates in the middle of the circuit, leaving gates near boundaries potentially trapped in barren plateau-like landscapes. This limitation is buried in the technical conditions rather than highlighted as a fundamental constraint.

The data loading contribution in Section 2 is not novel and not practically useful. The authors acknowledge that their circuits are classically simulable for fixed k , which means the “quantum” data loader offers no advantage over classical preprocessing. The suggestion that one could combine classically-trained HW-preserving blocks with harder-to-simulate components is pure speculation without theoretical or empirical support. How does trainability transfer across this transition? Does adding non-HW-preserving gates destroy the polynomial scaling? These questions are not addressed, making the data loading application an elaborate solution to a non-problem.

The comparison with prior work is misleading. The authors claim to show that Conjecture 1 from Larocca et al. “does not apply” to HW-preserving circuits, but this is because the conjecture’s hypotheses (full controllability, 2-design property) are not satisfied. This is not a refutation but a scope limitation. The positive contribution is providing an alternative analysis for circuits outside the conjecture’s domain, but the authors overstate this as revealing “a setting where the link between controllability and trainability does not apply.” In fact, there is still a link: lower controllability (smaller DLA dimension) correlates with better trainability (larger gradient variance), which is exactly what the DLA framework predicts for circuits that don’t achieve full controllability. The HW-preserving case is consistent with, not contradictory to, the broader theory.

Finally, the experimental validation is embarrassingly thin for a paper of this length. The Fashion MNIST experiments use a 5-qubit circuit with $k=2$, yielding a 10-dimensional encoding space, and report only average errors without statistical tests, baselines, or any connection to the theoretical results. The claim in Table 1 of “polynomial advantage” for quantum training is unsupported by any timing experiments. The code availability is not addressed, making the numerical claims unverifiable. This is not the empirical standard expected for publication in Quantum.

Verdict: 4/10. Recommendation: major-revisions bordering on reject.

6.2.5 Voice 5 — Editor-in-Chief synthesis

Having carefully considered all four reviews, I find substantial agreement on the paper’s strengths and weaknesses, with the Devil’s Advocate raising important concerns that sharpen the critique without fundamentally changing the assessment. The paper makes genuine theoretical contributions to understanding trainability of Hamming-weight preserving quantum circuits, but suffers from overselling of practical implications and inadequate empirical validation.

Reviewer 1 confirms the physics foundations are sound, with appropriate treatment of RBS/FBS gates and their relationship to fermionic systems. The concern about fermion-to-qubit mapping dependence (Jordan-Wigner vs. Bravyi-Kitaev) is technical but important for the completeness of the results. Reviewer 2’s critique of novelty is partially valid: the data loading application extends prior work incrementally, but the trainability analysis without 2-design assumptions does represent a methodological contribution. However, I agree with Reviewer 2 that the practical regime where fixed- k trainability guarantees are useful remains unclear. The fundamental tension between trainability (requiring small d_k) and expressivity (requiring large d_k) is acknowledged but not resolved, limiting the paper’s impact.

Reviewer 3 and the Devil’s Advocate both emphasize the inadequacy of empirical validation, and I concur this is a significant weakness. The reliance on Conjecture F1 without rigorous proof or comprehensive numerical support is concerning for a theoretical paper. The Fashion MNIST experiments are too superficial to demonstrate practical relevance. The authors must either strengthen the empirical evidence substantially or reframe the paper as purely theoretical, accepting that practical implications remain speculative. The Devil’s Advocate’s point about polynomial-but-still-impractical gradient scaling deserves explicit acknowledgment: showing $O(1/\text{poly}(n))$ instead of $O(1/\exp(n))$ is theoretically interesting but practically meaningful only if the polynomial degree permits optimization within reasonable shot budgets.

The comparison with the DLA-based framework is nuanced. The authors do not “refute” Conjecture 1 from Larocca et al. but rather demonstrate a complementary setting where alternative analysis techniques are needed. This should be reframed from confrontational to complementary positioning. The treatment of FBS gates’ reduced controllability and its implications for trainability is one of the paper’s genuine insights, showing that the controllability-trainability relationship is more subtle than simple inverse proportionality to DLA dimension would suggest.

Regarding the target journal’s requirements, the paper does include Author Contributions (Section 6) and Acknowledgments (Section 7), but lacks explicit Data Availability and Code Availability statements. The Methods content is distributed throughout the paper and appendices rather than consolidated, which is acceptable given the theoretical nature but should be rationalized. The paper fits Quantum’s scope and the detailed appendices are appropriate for this venue.

Must-fix items before resubmission (ordered by severity):

1. Explicitly quantify the practical regime where trainability guarantees are useful: for what (n, k) combinations does the polynomial gradient scaling permit optimization within feasible shot counts (e.g., 10 shots)?
2. Either prove Conjecture F1 or substantially expand numerical evidence to include diverse pattern architectures, larger qubit counts ($n \geq 100$), and characterization of the polynomial exponent as a function of circuit structure.
3. Add proper statistical analysis to gradient variance simulations: report chi-squared goodness-of-fit values, confidence intervals, and discuss any systematic deviations from theory.
4. Clarify the fermion-to-qubit mapping dependence of FBS results, specifically whether compound matrix relationships hold beyond Jordan-Wigner.
5. Add explicit Data Availability and Code Availability statements with repository links for reproducing all numerical results.
6. Reframe comparison with DLA framework from “refutation” to “complementary analysis,” accurately characterizing scope limitations rather than contradictions.
7. Either remove Fashion MNIST experiments or expand them to include proper baselines, statistical tests, and explicit connection to theoretical predictions.

6.2.6 Vote table

Voice	Recommendation	Confidence 1-10
Reviewer 1	minor-revisions	7
Reviewer 2	major-revisions	7
Reviewer 3	major-revisions	6
Devil’s Advocate	major-revisions	8
Editor-in-Chief	major-revisions	8

6.3 hwqml — Logical-fallacy report

Model: claude-opus-4-5-20251101 · 2,217 output tokens · 41.4s

Fallacy: hasty-generalization **Severity:** medium **Location:** Section 1 (Introduction), paragraph 2 — “Quantum machine learning (QML) has become a promising area for real world applications of quantum computers” **Why it’s the fallacy:** The abstract and introduction claim QML is “promising” for “real world applications” without citing evidence of demonstrated real-world utility. This generalizes from theoretical or small-scale demonstrations to broad practical applicability without sufficient supporting evidence. **Suggested fix:** Replace with a more measured claim: “Quantum machine learning (QML) has attracted significant research

interest as a potential application area for quantum computers, though practical demonstrations remain limited to small-scale problems.”

Fallacy: conflated-regimes **Severity:** medium **Location:** Section 3.2, Theorem 2 and surrounding discussion — “Thus, after some polynomial amount of repetitions, and for angles located at any constant fraction of the depth, there is an absence of Barren Plateaus for CPSA ansätze.” **Why it’s the fallacy:** The theorem requires L to grow “at least as fast as n^q ” (polynomial in qubit count), but the practical implications for scalability to large n are not rigorously addressed. The proof relies on Conjecture F1 about spectral gaps, with numerical evidence only provided for $n \in [4, 50]$ and $k = 1, 2, 3$. The claim of “absence of Barren Plateaus” is stated generally while the supporting evidence is limited to small system sizes. **Suggested fix:** Add explicit caveats: “For the system sizes tested numerically ($n \leq 50, k \leq 3$), we observe behavior consistent with absence of Barren Plateaus. Extension to larger systems relies on Conjecture F1, which remains unproven.”

Fallacy: asymptotic-only-claim **Severity:** medium **Location:** Section 3.2, Theorem 3 — “ $\text{Var}_{\theta} [C(\theta)] = k(n-k)/(n(n-1)) \hat{=} 8/d_k$ ” **Why it’s the fallacy:** The theorem establishes asymptotic scaling of gradient variance but does not provide finite- n bounds or discuss what values of n are sufficient for the asymptotic regime to accurately describe practical behavior. The verification plots (Figs. 9, 10) only show n up to 6 qubits. **Suggested fix:** Add a discussion of finite-size effects and the range of n for which the asymptotic formula provides accurate predictions, including error bounds.

Fallacy: active-space-handwave **Severity:** medium **Location:** Section 4.4 — “We do not claim that this plot exhibits any advantage to use RBS based quantum circuits as neural networks, but it illustrates that we can easily use such an architecture. Large simulation for more complex HW-preserving quantum neural networks for large value of k must be tackled in future work.” **Why it’s the fallacy:** The paper claims generalizability of the QNN approach to larger Hamming weights and more complex problems while explicitly acknowledging they have not actually run these experiments. This is a claim of generalization without supporting evidence. **Suggested fix:** Remove or weaken claims about generalizability to larger k until such experiments are conducted. The current disclaimer is appropriate but should be made more prominent.

Fallacy: circular-reasoning **Severity:** medium **Location:** Section 2.3, Algorithm 1 justification — “When the maximal rank (over parameter space) of the QFIM of a quantum data loader circuit is equal to $\dim(S^{\{d_k\}}) = d_k - 1$, we take it as evidence that it may achieve any state in $S^{\{d_k\}}$, i.e., achieve the amplitude encoding on the subspace of HW k .” **Why it’s the fallacy:** The paper uses QFIM rank as evidence of controllability, then uses controllability to justify the data loader design. However, high QFIM rank is a necessary but not sufficient condition for achieving arbitrary states—it indicates local controllability but does not prove global reachability. The reasoning is somewhat circular in using rank as both the design criterion and the validation. **Suggested fix:** Clarify that QFIM rank is a necessary condition that provides “support in favor of” but does not prove reachability. Add: “While maximal QFIM rank is necessary for full controllability, it is not sufficient; we therefore verify data loader capability empirically.”

Fallacy: appeal-to-authority **Severity:** medium **Location:** Section 5 (Discussion), paragraph 2 — “In recent work [11], authors have shown that if a subspace-preserving VQC satisfies the assumption of full controllability of the subspace... Our results are independent and consistent with these papers” **Why it’s the fallacy:** The paper repeatedly references the theoretical framework of [11, 17, 18] to position its results, but the relationship between these works and the current paper’s claims is not always clearly delineated. The appeal to these prior works is used to bolster credibility without fully establishing the logical connection. **Suggested fix:** More clearly delineate which results are proven independently versus which rely on the framework of prior work. Specify: “Our Theorem 2 and 3 are proven using direct analytical methods that do not invoke the DLA-based framework of [11, 17, 18].”

6.4 hwqml — Stage-6 CQE narrative

Model: claude-opus-4-5-20251101 · 1,121 output tokens · 31.9s

Process Summary: QuantumNovelty Run Evaluation

6.4.1 Composite Verdict

The run achieved a **composite score of 23 out of 100**, calculated via geometric mean across six dimensions. Per the SKILL.md scale interpretation, this places the collaboration firmly in the “**Inadequate**” tier (scores 20-39), indicating that while some structural elements were attempted, the run failed to produce the core artifacts required for a credible scientific contribution. A score of 23 suggests the process stalled early—likely during setup or initial exploration—before any substantive methodology could be executed.

This is not a borderline failure. The geometric mean’s sensitivity to low outliers means that multiple dimensions dragged the composite down, and indeed, the lowest-scoring dimension (Novelty rigour at 8) exerts disproportionate downward pressure. The run did not cross the threshold into “Marginally Acceptable” (40-59) because it lacks the fundamental outputs that would demonstrate even partial scientific validity.

6.4.2 Strongest Dimension: Communication (Score: 40)

Communication emerges as the relative high point, though “strongest” here is damning with faint praise. The score of 40 barely crosses into marginal acceptability. Both probes—**logical fallacies absent** and **reviewer panel verdict**—scored 40, but the evidence reveals why this dimension scored higher by default rather than by merit: the `logical_fallacies` skill was never run, and no `review_panel.md` was generated.

What this tells us about the run: the process never advanced far enough to produce text that could contain logical fallacies or receive reviewer scrutiny. The Communication score reflects an absence of failure rather than presence of quality. No claims were made poorly because no substantive claims were made at all. This is the scoring equivalent of a student receiving partial credit for writing their name on an otherwise blank exam.

6.4.3 Weakest Dimension: Novelty Rigour (Score: 8)

Novelty rigour scored a catastrophic 8 out of 100, anchoring the entire run in failure territory. The two probes reveal a complete breakdown in the baseline comparison stage:

- **Augmented baseline catalog present:** Scored 10/100 because “`baseline_catalog` has 0 rows.” This means the run failed to populate any prior work against which novelty could be measured. Without a baseline catalog, any claim of novelty would be unfounded—there’s nothing to demonstrate the work improves upon.
- **Strict-domination comparator run:** Scored 5/100 because “`novelty_verdict.json` not found.” The comparator that would formally establish whether proposed methods strictly dominate existing approaches never executed.

The stage that produced this failure is unambiguous: **literature synthesis and baseline establishment**. This is typically an early-phase activity, suggesting the run collapsed before meaningful research could begin. The absence of a populated baseline catalog indicates either a failure to retrieve relevant prior work, a breakdown in the catalog construction pipeline, or premature termination during the preparation phase.

A Novelty rigour score of 8 is not recoverable through downstream work—it represents a foundational gap that invalidates any subsequent findings.

6.4.4 Three Highest-Leverage Improvements

1. Prioritize Baseline Catalog Population Before Any Experimental Work The most critical fix is ensuring `baseline_catalog` is populated with non-zero rows before proceeding. This requires explicit checkpointing: the pipeline should halt and surface an error if the catalog remains empty after the literature synthesis stage. The next run should allocate dedicated resources to querying arxiv, semantic scholar, or domain-specific repositories for relevant prior work in the quantum computing space. Without baselines, novelty claims are scientifically meaningless.

2. Implement Artifact-Existence Gates for Stage Transitions Multiple probes failed because expected artifacts simply don't exist: `paper.tex` (False), `audit_claims.py` (False), `ablation_results.json` (False), `ratio_recompute.md` (False), `wilson_annotations.md` (False). The next run should enforce hard gates requiring each artifact's presence before advancing to subsequent stages. This prevents the cascade failure observed here, where early-stage incompleteness propagated through to universal artifact absence.

3. Execute Domain Specification Probes Early and Explicitly Domain depth scored 30 with all three probes—**active space stated explicitly**, **fermion-to-qubit mapping stated**, and **simulator precision floor disclosed**—failing due to absent references. These are foundational quantum chemistry specifications that should be declared in project setup, not discovered missing at evaluation time. The next run should include a mandatory domain specification template completed before any simulation work begins, ensuring active space definition, mapping choice (Jordan-Wigner, Bravyi-Kitaev, etc.), and precision constraints are documented upfront.

Bottom line: This run produced essentially no usable scientific output. The composite of 23 reflects systemic early-stage failure, not isolated weakness. Recovery requires rebuilding from baseline establishment forward, with artifact gates preventing silent failures from propagating.

7 qcnn — *Quantum Convolutional Neural Networks are Effectively Classically Simulable*

Pablo Bermejo, Paolo Braccia, Manuel S. Rudolph, Zoë Holmes, Lukasz Cincio, M. Cerezo [arXiv:2408.12739](https://arxiv.org/abs/2408.12739)
PRX Quantum 7, 020304 (2026)

7.1 qcnn — Deep-research review

Model: claude-haiku-4-5-20251001 · 1,328 output tokens · 37.5 s

7.1.1 1. One-paragraph summary of what the paper claims

The paper argues that Quantum Convolutional Neural Networks (QCNNS)—both tracing-out and measurement-based variants—are effectively classically simulable in a practical sense. The authors contend that randomly initialized QCNNS can only extract and process information encoded in low-bodyness (low-weight) Pauli observables of their input states, and that the benchmark datasets used in the literature to demonstrate QCNN success are “locally-easy,” meaning they can be classified using only this low-bodyness information. Based on these insights, the authors construct purely classical surrogates using Pauli propagation methods (LOWESA), tensor networks, and classical shadows that match or outperform standard QCNNS on all tested benchmark datasets at scales up to 1024 qubits. They conclude that there is currently no evidence QCNNS will work on classically non-trivial tasks, and challenge the community to identify datasets where QCNNS provide genuine quantum advantage.

7.1.2 2. Audit-and-falsify checklist

Item	Status	Evidence
Augmented baseline catalog	PARTIAL	The paper compares against “full QCNNS” trained without finite sampling and references prior literature results (e.g., Table I cites accuracies from Refs. [51–69]), but does not systematically benchmark against other recent classical ML baselines (random forests, CNNs, kernel methods) on the same datasets beyond the single random-forest demonstration in Appendix D.
Strict-domination comparator	FAIL	Claims of “matching or outperforming” are made at displayed precision (e.g., “98% vs. 98%”) without specifying tolerance thresholds (<code>_abs</code> , <code>_rel</code>); no formal Pareto analysis with calibrated tolerances is provided.
Recompute-from-raw	PARTIAL	Table I displays test accuracies directly, but no raw confusion matrices, per-run values, or derivation scripts are presented; readers cannot independently verify that “best out of 5” selections are consistent with underlying distributions.
Wilson 95% CIs	FAIL	Small-sample classification results (e.g., 5 independent runs, 100 test points) are reported without binomial confidence intervals; statements like “100/98” accuracy lack uncertainty quantification.
Cross-LLM falsifiability	NOT-APPLICABLE	No LLM-in-the-loop method is employed in this work.

Item	Status	Evidence
Honest negatives	PARTIAL	The authors acknowledge that their results do not constitute full dequantization and note caveats (e.g., 2D QCNNs may be harder to simulate, misclassifications near phase boundaries), but there is no dedicated “Failure Modes” section cataloging cases where the classical surrogate underperformed or failed to converge.
Simulator precision floor	FAIL	No discussion of numerical precision (float32 vs. float64) in energy or expectation-value comparisons; the DMRG states and shadow tomography reconstructions are not validated against a float64 reference path.
Auditable claims	FAIL	No <code>audit_claims.py</code> or equivalent re-runnable script is referenced; data files (JSON, HDF5) and code are not explicitly linked or described as available for reproducing every numerical claim from raw artifacts.

7.1.3 3. Overall assessment

This paper presents a conceptually significant and technically sophisticated argument that QCNNs’ heuristic success may be attributable to classically simulable dynamics on locally-easy datasets. The theoretical underpinnings (Result 1, Result 2, Theorem 1) are carefully developed using Weingarten calculus, and the empirical demonstrations span multiple quantum and classical datasets at impressive scales. However, the paper falls short of the research rigor expected by a strict audit framework: quantitative claims lack uncertainty estimates, baseline comparisons are incomplete, numerical precision is unaddressed, and reproducibility infrastructure (auditable scripts, raw data archives) is absent. The “best of 5” reporting and absence of confidence intervals on small-sample accuracies weaken the statistical credibility of the comparative claims. While the core thesis is compelling and the simulations are impressive, the evidentiary standards do not meet those required for a fully auditable, falsifiable research contribution.

Rigour score: 5/10

7.1.4 4. Three highest-leverage improvements

- 1. Add Wilson 95% confidence intervals to all reported accuracies.** For every entry in Table I and Figures 3–6, compute and display binomial CIs given the test-set size. Replace statements like “98%” with “98% [92.9–99.8%, Wilson 95% CI, n=100].” This single change would substantially increase statistical credibility.
- 2. Provide an `audit_claims.py` script with on-disk JSON artifacts.** Archive all raw experimental outputs (per-run accuracies, shadow samples, trained parameters) in a public repository, and include a single script that re-derives every numerical claim (table entries, figure data points, ratios) from these artifacts. This converts the paper from “trust-based” to “verify-based” reproducibility.
- 3. Establish a float64 reference path for simulator precision.** Re-run at least one representative quantum-dataset experiment (e.g., the 1024-qubit XXX model) with float64 throughout the DMRG, shadow tomography, and LOWESA pipelines; compare against the default (likely float32/complex64) results and report the deviation. If deviations are negligible, state this explicitly; if not, adjust conclusions accordingly.

7.2 qcnn — Reviewer panel 5 voices

7.2.1 Voice 1 — Reviewer 1 (Physics correctness)

This manuscript addresses an important question in quantum machine learning: whether the heuristic success of quantum convolutional neural networks can be attributed to their operation within a classically simulable regime. The authors present a coherent theoretical framework connecting barren plateau avoidance to classical simulability via low-bodyness operator subspaces, and they provide extensive numerical demonstrations. From a physics correctness standpoint, the core theoretical claims rest on well-established Weingarten calculus techniques and the structure of unitary t -designs, which are applied appropriately here.

The treatment of the Heisenberg-evolved measurement operator and its decomposition into k -purity contributions (Equation A1 and following) is mathematically sound. Result 1 and Result 2 are stated informally in the main text but formalized properly in Theorem 1 of the appendix, where the authors derive exact expressions for the average k -purities of a prototypical QCNN ansatz. The key insight that randomly initialized QCNNs with 2-design convolutional gates predominantly support low-bodyness observables follows from the structure of the P -gate (Equation A11), which projects onto the commutant of $U(4)^2$. The exponential decay of contributions from high-bodyness Paulis with increasing bodyness is demonstrated both analytically and through Figure 8, which shows the distribution for $n=1024$ qubits. However, I note that the prefactor $2/5$ appearing in the P -gate originates from the 15-dimensional Weingarten matrix for $U(4)$, and while the authors cite appropriate references, an explicit derivation in the supplementary material would strengthen reproducibility.

The physics of the condensed matter Hamiltonians used as quantum datasets appears correct. The bond-alternating XXX model (Equation 8), Haldane chain (Equation 9), ANNNI model (Equation 10), and cluster Hamiltonian (Equation 11) are standard models with well-characterized phase diagrams. The authors use DMRG to obtain ground states, which is appropriate for these one-dimensional systems. However, I have concerns about the phase boundaries employed. For instance, the Haldane chain phase transition is quoted as occurring at $h=0.423$ for fixed $h=0.5$ and $J=1$, attributed to thermodynamic limit analysis. For finite-size systems of $n=512$ qubits, finite-size corrections to the critical point could be significant, and the authors acknowledge that misclassifications near phase boundaries may arise from using thermodynamic limit labels. This is a reasonable caveat, but a more quantitative analysis of finite-size effects would strengthen the claims about classification accuracy.

The connection between entanglement structure and QCNN trainability deserves further attention. The authors correctly note that volume-law entangled states can suppress the low-bodyness contributions that QCNNs rely upon, leading to barren plateaus. However, the quantum datasets considered (ground states of gapped local Hamiltonians in one dimension) are known to satisfy area-law entanglement, which naturally admits efficient MPS representations. This raises the question of whether the simulability demonstrated here is primarily a consequence of the data structure rather than the QCNN architecture itself. The tensor network simulations in Section IV.B and Appendix E confirm that the input states admit efficient classical representations, but this conflates two separate sources of simulability.

Questions for Authors: First, can you provide tighter bounds on the finite-size corrections to phase boundaries for the quantum datasets, and how do these affect your reported classification accuracies? Second, for the measurement-based QCNN analysis in Appendix E, you show bond dimension scaling as $\sim n/8$, but this appears to be an empirical observation from random unitaries. Can you provide analytical bounds or worst-case guarantees? Third, the paper claims QCNNs “cannot leave” the low-bodyness subspace during training, but the theoretical results (Result 1, Result 2) are average-case statements. What prevents the optimizer from finding parameter regions where high-bodyness contributions become significant?

Verdict: 7/10. Recommendation: minor-revisions.

7.2.2 Voice 2 — Reviewer 2 (Algorithmic novelty)

The central algorithmic contribution of this work is the construction of classical surrogates for QCNNs using low-bodyness Pauli propagation and tensor network methods, demonstrating that these surrogates can match or exceed the classification performance of actual QCNNs. Evaluating this against the recent literature requires careful consideration of what has been established in the past 24 months regarding classical simulation of variational quantum circuits.

The most directly relevant prior work is Reference 49 (Cerezo et al., Nature Communications 2025), which established the conceptual link between barren plateau absence and classical simulability but did not provide end-to-end training demonstrations. Reference 88 (Angrisani et al., PRL 2025) showed average-case simulability of noiseless circuits via low-weight Pauli truncation but similarly did not demonstrate successful training over

simulated landscapes. The present work’s primary novelty lies in closing this gap: showing that one can not only estimate loss functions at random parameter points but also successfully train a classical surrogate to solve the same tasks that QCNNS purportedly solve. This is a meaningful advance, though the authors are appropriately careful to note that this constitutes a “proof by demonstration” rather than a rigorous worst-case guarantee.

Compared to Schreiber, Eisert, and Meyer (PRL 2023, Reference 131) on classical surrogates for quantum learning models, the present work specifically targets the QCNN architecture and leverages its structural properties (logarithmic depth, local pooling) rather than general variational circuits. The LOWESA algorithm employed here (References 89, 90) provides the Pauli propagation backbone, but the authors introduce novel truncation strategies including variance-based operator selection and sliding-window locality restrictions. The tensor network approach in Appendix B2 with constrained bodyness is also technically novel, though it builds on the MPS projection methods of Reference 114. However, I find the novelty somewhat incremental: the theoretical insights largely follow from combining known results about QCNN structure with established Pauli propagation techniques.

The empirical comparisons present a Pareto-dominance argument: the classical surrogates achieve comparable or better accuracy while requiring dramatically fewer quantum resources. For instance, the authors claim that for the XXX model at $n=1024$ qubits, successful classification is achieved with only 100 classical shadows per data point and 20,000 total measurement shots, compared to the 5,000-10,000 shots per iteration per data point required for standard QCNN training. These ratios appear recomputable from the stated parameters, though I could not independently verify them without access to the raw experimental data. The classical simulation resources (stated to be achievable on a modern laptop) are not precisely quantified in terms of wall-clock time or memory usage, which makes the Pareto comparison somewhat informal.

A significant concern regarding algorithmic novelty is the scope of applicability. The authors focus exclusively on one-dimensional QCNNS with nearest-neighbor connectivity. They acknowledge that two-dimensional or more exotic topologies would increase simulation costs, potentially prohibitively. They cite Reference 124 (Napp et al., PRX 2022) on 2D random circuit simulation but do not attempt such simulations. Given that practical applications of QCNNS might require higher-dimensional architectures (e.g., for image classification with spatial structure), the restriction to 1D significantly limits the impact of the results. The claim that “our community is in dire need of non-trivial datasets” is provocative but somewhat deflects from addressing whether the simulability results extend to more complex architectures.

Questions for Authors: First, can you provide explicit wall-clock time and memory comparisons between your classical surrogates and actual QCNN implementations on equivalent hardware? Second, have you attempted any simulations of 2D QCNNS, and if so, at what system sizes does the simulation cost become prohibitive? Third, the variance-based operator selection heuristic (keeping only operators with high variance across the dataset) is described briefly. How sensitive are the results to the threshold used, and is there theoretical justification for this approach?

Verdict: 6/10. Recommendation: minor-revisions.

7.2.3 Voice 3 — Reviewer 3 (Empirical evidence)

The empirical claims in this manuscript span both quantum and classical datasets across multiple system sizes, requiring careful evaluation of statistical rigor, reproducibility, and completeness of ablations. The authors present classification accuracies for eight distinct tasks, with system sizes ranging from $n=32$ to $n=1024$ qubits, making this among the largest-scale QCNN simulation studies to date.

Regarding statistical confidence intervals, the reporting is inconsistent throughout the paper. For the quantum datasets, the authors state that training is “averaged over 5 different runs” but do not report standard deviations or confidence intervals on the classification accuracies. For instance, the XXX model results (Figure 3b) show test accuracies above 90% at 100 shadows per data point, but without variance information it is impossible to assess the reliability of this claim. A Wilson 95% confidence interval for a classification rate of 90% on a test set of 25 samples (as implied by the 100-sample dataset with 75 training points) would span approximately [72%, 97%], which is quite wide. The multi-class classification results for ANNNI and Cluster models report single-point accuracies (e.g., “train accuracy of 82.8%, test accuracy of 85.8%”) without uncertainty quantification. Table I presents “best out of 5 independent runs” for classical datasets, which is methodologically problematic as it cherry-picks favorable results rather than reporting mean performance.

The ablation study coverage is incomplete. The authors vary the number of training points and number of shadows per data point (Figures 3, 4, 10, 11), which constitutes a sensitivity analysis for the shadow tomography

component. However, there is no systematic ablation of the bodyness truncation threshold. The text mentions using “maximum bodyness of two” for XXX, “three” for Haldane, and “four” for ANNNI and Cluster, but there is no justification for these choices or exploration of how performance degrades if the threshold is set too low. Similarly, the frequency truncation in the LOWESA algorithm is mentioned but not systematically varied. The sliding-window heuristic for operator selection (restricting non-identity Paulis to adjacent qubits) is applied but not ablated against a full-operator baseline.

The absence of a dedicated failure-modes section is notable. While the authors acknowledge that QCNNS could potentially succeed on non-locally-easy datasets, they do not present any examples where their classical simulation fails to match QCNN performance. This creates an unfalsifiable narrative: if the classical surrogate succeeds, the dataset is declared “locally easy,” and if the QCNN itself fails, the dataset is too entangled. A more rigorous approach would construct synthetic datasets with tunable complexity (e.g., by varying entanglement structure) and demonstrate a transition from simulable to non-simulable regimes. The random forest analysis in Appendix D provides some evidence that local observables suffice for classification, but this is only shown for a single dataset (XXX model).

Regarding reproducibility, the authors do not mention whether code or data will be made publicly available. The implementation details are spread across the main text and four appendices, but critical parameters (learning rates, number of optimizer iterations, LBFGS hyperparameters) are not fully specified. The DMRG simulations use ITensor.jl, but convergence criteria and bond dimension cutoffs are not reported. For the classical datasets, the image preprocessing pipeline (resizing, grayscale conversion, MPS encoding) is described qualitatively but not with sufficient detail for exact reproduction. The absence of an audit script or supplementary data files that would allow independent verification of the numerical claims is a significant weakness for a paper making strong claims about computational resources.

Questions for Authors: First, can you provide standard deviations across the 5 runs for all reported classification accuracies, and compute proper confidence intervals given your test set sizes? Second, what happens to classification accuracy if you reduce the bodyness truncation threshold by one for each dataset? Third, will code and data be released upon publication, and if so, will this include the raw shadow measurement data and trained parameters?

Verdict: 5/10. Recommendation: major-revisions.

7.2.4 Voice 4 — Devil’s Advocate

This paper’s central claim, that QCNNS are “effectively classically simulable,” rests on a foundation of circular reasoning that the other reviewers have been too generous in overlooking. The argument proceeds as follows: QCNNS succeed only on datasets that are classifiable via low-bodyness observables; the authors call such datasets “locally easy” (Definition 1); therefore, QCNNS are classically simulable on the datasets where they succeed. But this is not a discovery about QCNNS; it is a tautology dressed in Weingarten calculus. The authors have not shown that QCNNS cannot succeed on non-locally-easy datasets, only that they have not been tested on such datasets. The burden of proof they claim to shift onto QCNN proponents could equally be shifted onto them: construct a locally-hard dataset and show that QCNNS fail on it.

The theoretical results (Result 1, Result 2, Theorem 1) are average-case statements that do not constrain the behavior of trained QCNNS. The authors acknowledge this repeatedly but then proceed to draw conclusions as if it were irrelevant. The claim that “the QCNN’s training process is initially guided by low-bodyness information” and that “if the landscape is sufficiently well-behaved... then the QCNN could accurately solve the task at hand” is pure speculation. The Pauli propagation surrogate faithfully simulates the QCNN only if the QCNN never leaves the low-bodyness subspace during training, but there is no theorem guaranteeing this. The numerical demonstrations show that surrogates work on specific datasets, but these are the same datasets where QCNNS were known to succeed, creating a confirmation bias. If I construct a dataset where the QCNN succeeds but the surrogate fails, does that disprove the paper? The authors provide no criteria for what would constitute a refutation.

The claim of “substantially reduced quantum resources” is misleading. The authors compare their shadow-based approach to a strawman implementation of QCNN training that uses 5,000-10,000 shots per iteration per data point. But modern variational quantum algorithms routinely use variance-reduced gradient estimators, parameter-shift rules with measurement reuse, and other optimizations that dramatically reduce shot counts. Reference 12-24 in the paper discuss these trainability issues, but the resource comparison ignores decades of progress in efficient gradient estimation. Furthermore, the classical simulation cost is never honestly reported.

The authors state their simulations run “on a modern laptop” but do not specify CPU time, memory usage, or how these scale with system size. For the $n=1024$ qubit XXX model, the LOWESA algorithm must track an exponentially growing number of Pauli paths before truncation. The frequency truncation prevents this from exploding, but at what cost to fidelity? The authors do not report the approximation error introduced by their truncations.

The empirical methodology has serious flaws beyond those identified by Reviewer 3. The selection of “400 operators with the largest variance across the shadows dataset” (Section IV.A.1) is a form of data snooping: using the test data to select features before training. This inflates classification accuracy and makes the comparison to QCNNS unfair. The authors apply this heuristic without acknowledgment that it constitutes a methodological violation. Similarly, the use of “best out of 5 runs” for classical dataset results (Table I) is cherry-picking that would not survive peer review at a machine learning venue. The phase diagram classifications (Figures 5, 6) show systematic errors near phase boundaries that the authors attribute to finite-size effects, but these could equally indicate that the low-bodyness approximation breaks down precisely where classification is hardest. The 77-85% test accuracies reported for ANNNI and Cluster models are not competitive with state-of-the-art classical methods for phase classification, undermining the claim that classical shadows plus LOWESA constitute a practical alternative.

The paper’s rhetoric is designed to shift burden of proof in a way that is not scientifically productive. Statements like “the burden of proof now rests firmly in the hands of any proponent of QCNNS” and “we boldly claim: There is currently no evidence that QCNNS will work on classically non-trivial tasks” are advocacy, not science. The authors have shown that specific implementations of QCNNS on specific datasets can be simulated, not that QCNNS as a model class are fundamentally limited. The restriction to one-dimensional architectures with nearest-neighbor connectivity is severe, and the hand-waving about “advanced tensor network techniques” for higher-dimensional cases (Section V) does not constitute a result. This paper would have been much stronger as a careful case study of 1D QCNN simulability rather than a sweeping indictment of an entire research direction.

Recommendation: major-revisions, verging on reject. The paper makes important technical contributions but wraps them in unsupported rhetorical claims that undermine scientific credibility.

7.2.5 Voice 5 — Editor-in-Chief synthesis

Having reviewed all four assessments, I find substantial agreement on the technical merit of the core contributions but significant disagreement on whether the claims are appropriately scoped and supported. Reviewer 1 finds the physics correct with minor concerns about finite-size effects and the conflation of data simulability with architecture simulability. Reviewer 2 acknowledges incremental but meaningful algorithmic novelty while noting the restriction to 1D architectures. Reviewer 3 raises serious concerns about statistical reporting and ablation completeness. The Devil’s Advocate argues that the paper’s central thesis is circular and its rhetoric oversteps the evidence.

I find the Devil’s Advocate’s critique of circularity partially valid but overstated. The authors do provide value by demonstrating explicitly that standard QCNN benchmarks are classically simulable, even if this could be anticipated from theoretical considerations. The practical demonstration matters for the field. However, the Advocate’s point about burden-shifting rhetoric is well-taken: the provocative framing (“we boldly claim”) is inappropriate for a scientific publication and should be moderated. The paper should present its findings as characterizing the simulability regime of specific QCNN implementations rather than dismissing QCNNS categorically.

The statistical deficiencies identified by Reviewer 3 are disqualifying in their current form. Reporting “best out of 5 runs” without mean and variance is unacceptable for claims about comparative performance. The confidence interval issue is particularly acute for the small test sets used: classification accuracies must be reported with proper uncertainty quantification. The variance-based operator selection (using test data characteristics to select features) is methodologically problematic and must be addressed either by using only training data for selection or by clearly acknowledging the information leakage. These issues can be corrected with additional analysis of existing data.

The scope limitations (1D architectures only, ground states of gapped Hamiltonians only) should be stated more prominently as caveats rather than buried in the discussion. The paper’s contribution is significant for the specific regime studied, but the broad claims about QCNNS being “classically simulable” need qualification. PRX Quantum’s standards require that methodological limitations be clearly articulated, and the current presentation does not meet this bar.

My decision is major-revisions. The technical core is sound and the demonstration is valuable, but the presentation requires significant modification to bring rhetorical claims in line with evidence, and the statistical analysis must be brought to publishable standards. The following must-fix items are ordered by severity:

1. Replace all “best out of 5” results with mean and standard deviation across runs; compute and report Wilson confidence intervals for all classification accuracies given test set sizes.
2. Address the feature selection methodology: either restrict operator variance calculation to training data only, or acknowledge and quantify the information leakage from test data.
3. Moderate rhetorical claims throughout: replace “we boldly claim” framing with neutral scientific language; explicitly state that results apply to 1D QCNNS on area-law entangled data.
4. Add an ablation of bodyness truncation threshold showing how classification accuracy degrades as the threshold is reduced.
5. Report explicit computational costs (wall-clock time, memory, scaling) for classical simulations to enable fair comparison with actual QCNN implementations.
6. Provide code and data availability statement; ideally release shadow data and trained parameters for reproducibility.
7. Address finite-size effects on phase boundaries quantitatively, or remove claims about classification accuracy being meaningful near phase transitions.

7.2.6 Vote table

Voice	Recommendation	Confidence 1-10
Reviewer 1	minor-revisions	7
Reviewer 2	minor-revisions	6
Reviewer 3	major-revisions	8
Devil’s Advocate	major-revisions	7
Editor-in-Chief	major-revisions	8

7.3 qcnn — Logical-fallacy report

Model: claude-haiku-4-5-20251001 · 2,494 output tokens · 44.5 s

Fallacy: cherry-picked-baseline **Severity:** medium **Location:** Section IV.A, paragraph introducing quantum datasets: “Notably, the classical, weight-truncated QCNN is self-consistently trained to solve the classification tasks, not to faithfully emulate the training of an exact QCNN.” **Why it’s the fallacy:** The authors explicitly state they are not comparing their classical surrogate against a faithfully trained QCNN but rather training their classical model directly on the task. This setup avoids the harder comparison of whether their surrogate can match a QCNN that has been optimally trained with full quantum resources. By training the surrogate “self-consistently” rather than benchmarking against the best possible QCNN performance, they sidestep a stronger baseline comparison. **Suggested fix:** Include explicit comparisons where a QCNN is trained with sufficient shots and optimization iterations to reach its best achievable accuracy, then compare the classical surrogate’s accuracy against this optimized QCNN performance on the same datasets.

Fallacy: conflated-regimes **Severity:** medium **Location:** Section V, Discussion: “While we focus here on the two most popular instantiations of QCNNS used in the literature (one-dimensional tracing out and measurement-based architectures), it is clear that these are the easiest to classically simulate. One could, for instance, envision QCNNS in two or more dimensions, as well as more exotic topologies.” **Why it’s the fallacy:** The authors acknowledge their results apply specifically to one-dimensional QCNNS but then make broad claims throughout the paper about QCNN simulability in general. The title “Quantum Convolutional

Neural Networks are Effectively Classically Simulable” and bold claims like “There is currently no evidence that QCNNS will work on classically non-trivial tasks” extrapolate from the 1D case to all QCNNS without demonstrating results for higher-dimensional or more complex topologies. **Suggested fix:** Modify the title to “One-Dimensional Quantum Convolutional Neural Networks are Effectively Classically Simulable” and qualify all general statements about QCNNS to explicitly state they apply only to the 1D architectures studied.

Fallacy: active-space-handwave **Severity:** medium **Location:** Section V, Discussion: “We strongly believe that the techniques introduced here can serve as blueprints to classically simulate other architectures.” **Why it’s the fallacy:** The authors claim their techniques generalize to other quantum neural network architectures without actually running experiments or providing rigorous proofs for these other architectures. The phrase “we strongly believe” signals speculation rather than demonstrated results. This constitutes handwaving about generalization capability without empirical validation. **Suggested fix:** Either remove claims about generalization to other architectures or include explicit experimental results demonstrating classical simulation of at least one additional architecture type beyond QCNNS.

Fallacy: hasty-generalization **Severity:** medium **Location:** Section IV.B, Classical datasets conclusion: “Concomitantly, this implies that using QCNN-based QML schemes for classical data appears to be an ill-motivated task.” **Why it’s the fallacy:** The authors tested only four classical datasets (MNIST, Fashion-MNIST, EuroSAT, GTSRB) with specific encoding schemes and conclude that QCNNS are “ill-motivated” for all classical data. This sweeping conclusion is drawn from a limited sample that does not represent the full diversity of classical data classification problems or encoding strategies. **Suggested fix:** Qualify the conclusion to state: “For the classical datasets and encoding schemes tested in this work, QCNN-based approaches appear to offer no advantage over classical simulation. Whether this extends to all classical data problems remains an open question.”

Fallacy: cherry-picked-baseline **Severity:** medium **Location:** Section IV.A.3, ANNNI model: “Thus we can reach similar results to those obtained in the literature, at a much smaller measurement budget.” **Why it’s the fallacy:** The comparison claims to match “results obtained in the literature” but does not specify which literature results are being compared against, what their experimental conditions were, or whether those prior results used optimized QCNN configurations. Without explicit citations and controlled comparisons, this claim of matching performance lacks rigor. **Suggested fix:** Add explicit citations to the specific prior QCNN results being compared against, including their reported accuracies, number of measurements used, and experimental conditions, then present a direct numerical comparison table.

Fallacy: asymptotic-only-claim **Severity:** medium **Location:** Section IV.A.1, XXX model: “To our knowledge, this constitutes the largest QCNN implementation with classical shadows” **Why it’s the fallacy:** While the authors demonstrate 1024 qubits, they use this finite-N demonstration to support broader claims about QCNN simulability without acknowledging potential scaling issues. The complexity analysis in Appendix E showing $\sim n^8$ bond dimension scaling means simulation cost scales as $O(n^{24})$, which could become intractable at larger scales despite the polynomial label. **Suggested fix:** Include explicit discussion of the scaling limitations, noting that while $n=1024$ is achievable, the $O(n^{24})$ cost scaling means practical limits exist, and specify the largest system size that remains tractable on available hardware.

Fallacy: pareto-cherry-picked-axes **Severity:** medium **Location:** Table I and Section IV.B: “In Table I we show results for all the classification tasks considered. Here we can see that the simulated QCNN result in high test accuracies (showing best out of 5 independent runs), comparable to, and even larger than, those found in the literature.” **Why it’s the fallacy:** By reporting “best out of 5 independent runs” rather than mean \pm standard deviation, the authors cherry-pick the most favorable outcome along the accuracy axis. This selection bias inflates apparent performance by ignoring variability across runs. **Suggested fix:** Report mean accuracy \pm standard deviation across all 5 runs for each dataset and encoding combination, not just the best single run.

7.4 qcnn — Stage-6 CQE narrative

Model: claude-opus-4-5-20251101 · 1,142 output tokens · 29.7 s

Process Summary: QuantumNovelty Run Evaluation

7.4.1 Composite Verdict

The geometric mean composite score of **23 out of 100** places this run firmly in the **“Inadequate”** tier of collaboration quality. Per the SKILL.md interpretation guidelines, scores below 30 indicate fundamental process failures that undermine the credibility of any claims the run might produce. This is not a borderline result requiring nuanced interpretation—the run failed to execute core methodological stages that distinguish rigorous research from exploratory prototyping.

A geometric mean of 23 derived from six dimensions ranging from 8 to 40 reveals no dimension achieved even passable quality. The geometric mean’s sensitivity to low outliers correctly penalizes the run: a single catastrophic dimension (Novelty rigour at 8) drags the composite below what an arithmetic mean would suggest. This is appropriate—research quality is multiplicative, not additive. Weak novelty verification invalidates downstream claims regardless of how well other stages performed.

7.4.2 Strongest Dimension: Communication (Score: 40)

Communication emerged as the strongest dimension, though “strongest” here means “least deficient” rather than “adequate.” Both probes—“logical fallacies absent” and “reviewer panel verdict”—scored 40, but critically, this score reflects *absence of evaluation* rather than confirmed quality. The evidence for both probes indicates the relevant skill modules were never executed: “logical_fallacies skill not run” and “no review_panel.md found.”

This pattern reveals something important about the run’s failure mode. Communication scored highest not because the run communicated well, but because it never progressed far enough to produce artifacts that could be evaluated for communication quality. The run collapsed at earlier stages—reproducibility, methodology, domain depth—before generating sufficient output for communication assessment. A score of 40 based on unevaluated criteria is epistemically hollow; it tells us nothing about actual communication quality and everything about premature termination.

7.4.3 Weakest Dimension: Novelty Rigour (Score: 8)

Novelty rigour scored 8, the lowest of any dimension, indicating near-total failure of the novelty verification stage. The probe breakdown is damning:

- **“augmented baseline catalog present”** scored 10 with evidence showing “baseline_catalog has 0 rows.” The run produced an empty catalog structure—the scaffolding existed but contained no actual baseline comparisons. This suggests the catalog generation code executed but either received no input data or encountered silent failures that produced empty output.
- **“strict-domination comparator run”** scored 5 with evidence “novelty_verdict.json not found.” The comparator stage never executed or failed without producing output. Without a novelty verdict artifact, any claims of novel contribution are unsupported assertions.

The stage that produced this failure is unambiguously the **baseline cataloging and comparison stage**. Zero baseline rows means no comparison targets existed when the strict-domination comparator attempted to run. This is a dependency cascade: baseline cataloging failed first (producing 0 rows), which made comparison impossible (no verdict file). The root cause lies in whatever process should have populated the baseline catalog—likely a literature ingestion or prior-work extraction module that either wasn’t invoked or failed silently.

7.4.4 Three Highest-Leverage Improvements

1. Implement Baseline Catalog Validation Gates The empty baseline catalog should have halted the pipeline immediately. The run continued executing downstream stages despite having no comparison targets, wasting computation and producing meaningless outputs. **Concrete fix:** Add a hard gate after baseline

cataloging that requires `baseline_catalog.rows >= N` (where N is a domain-appropriate minimum, likely 5-10 for quantum chemistry). Pipeline stages beyond this gate should refuse to execute if the predicate fails. This prevents the cascade of wasted work visible in this run.

2. Emit Core Artifacts Before Optional Analyses Multiple probes across Reproducibility and Methodological rigour dimensions failed because expected artifacts don't exist: `audit_claims.py`, `paper.tex`, `wilson_annotations.m`, `ablation_results.json`, `ratio_recompute.md`. The Pareto archive exists but contains 0 rows. **Concrete fix:** Restructure the pipeline to emit minimal viable versions of core artifacts early, then enrich them. A skeleton `paper.tex` with section headers should exist before any analysis runs. This ensures evaluation probes have targets and surfaces structural failures earlier.

3. Add Vendor Configuration Verification The Falsifiability probe “cross-LLM with multiple vendors” shows `vendors used: []`—an empty list. A run claiming to evaluate LLM collaboration must actually invoke LLM vendors. **Concrete fix:** Add a pre-flight check that verifies vendor configuration is non-empty and that at least one vendor responds successfully before the main run begins. This catches configuration errors, credential issues, and network problems before they waste an entire run.

This run represents a process failure, not a research failure. The methodology exists but wasn't executed. Priority one is ensuring the pipeline cannot progress past failed dependencies.

8 majorana2 — 20 Second Parity Lifetime in an InAs-Pb Tetron Device

Microsoft Quantum [arXiv:2606.03884](https://arxiv.org/abs/2606.03884) [arXiv:2606.03884](https://arxiv.org/abs/2606.03884) (preprint, June 2026)

8.1 majorana2 — Deep-research review

Model: claude-opus-4-5-20251101 · 1,451 output tokens · 39.8 s

8.1.1 1. One-paragraph summary of what the paper claims

This paper from Microsoft Quantum reports a major advance in topological quantum computing by demonstrating a characteristic parity lifetime of approximately 20 seconds in an InAs-Pb tetron device, representing an improvement of more than three orders of magnitude over previous Al-based devices (which achieved 1–12 ms). The authors attribute this improvement to replacing aluminum with the higher-gap superconductor lead (Pb) in their superconductor-semiconductor hybrid platform, which yields a larger topological gap ($\sim 70 \text{ } \mu\text{eV}$ vs. $\sim 30 \text{ } \mu\text{eV}$ in Al devices) and a proximity-induced gap of $\sim 570 \text{ } \mu\text{eV}$. They introduce an rf-based wire spectroscopy technique for scalable device bring-up that can resolve Majorana hybridization energies (EM) with μeV precision, and demonstrate $h/2e$ -periodic bimodal quantum capacitance shifts consistent with parity-dependent interference. The paper argues this validates a central premise of topological quantum computing: that increasing the excitation gap dramatically reduces error mechanisms and improves qubit performance.

8.1.2 2. Audit-and-falsify checklist

Item	Status	Evidence
Augmented baseline catalog	PASS	The paper explicitly compares against their own prior Al-InAs devices (Refs. 59–61) with quantitative improvements cited: parity lifetime improvement from 1–12 ms to ~ 20 s; topological gap from $\sim 30 \text{ } \mu\text{eV}$ to $\sim 70 \text{ } \mu\text{eV}$; topological phase region more than doubled; this represents comparison against current state-of-the-art from the same research group.
Strict-domination comparator	PARTIAL	The paper reports “more than three orders of magnitude” improvement in parity lifetime and provides specific values (22 ± 1 s fit), but comparisons like “top quintile gap” thresholds and localization length claims lack explicit tolerance specifications (<code>_abs</code> , <code>_rel</code>).
Recompute-from-raw	PARTIAL	The paper presents raw data in figures (e.g., time traces in Fig. 6, phase diagrams in Fig. 3) and reports derived values ($Z = 22 \pm 1$ s from $N = 324$ dwell intervals), but there is no explicit statement that readers can access raw data or re-derive the numerical claims independently.
Wilson 95% CIs	PARTIAL	The parity lifetime $Z = 22 \pm 1$ s includes uncertainty from exponential fitting, but the paper does not explicitly provide binomial/Wilson confidence intervals for small-sample statistics; the $N = 324$ events is reasonably large but the methodology for uncertainty propagation is not fully detailed.
Cross-LLM falsifiability	NOT-APPLICABLE	No LLM-in-the-loop methods were used in this experimental physics paper.

Item	Status	Evidence
Honest negatives	PARTIAL	The paper acknowledges that “external quasiparticle poisoning events” cause sign switches in their data (Sec. 3–4), and discusses that non-equilibrium quasiparticles can degrade protection, but there is no dedicated “Failure Modes” section systematically cataloging regimes where the device underperforms or the method fails.
Simulator precision floor	PARTIAL	The paper mentions simulations for induced gap values (“consistent with simulations”) and uses numerical modeling, but does not explicitly state whether float64 reference paths were used or address numerical precision floors in their calculations.
Auditable claims	FAIL	There is no mention of a re-runnable script (e.g., <code>audit_claims.py</code>), on-disk JSON, or any reproducibility infrastructure that would allow independent derivation of numerical claims from raw data.

8.1.3 3. Overall assessment

This paper represents solid experimental physics work with clear advances over prior results from the same group. The claims are well-grounded in extensive measurements, the methodology (rf-based spectroscopy, parity injection) is clearly described, and the results are internally consistent. However, from a strict research-rigour audit perspective, the paper falls short in several areas: (1) no publicly accessible raw data or reproducibility infrastructure; (2) uncertainty quantification is present but not comprehensively specified (e.g., explicit confidence interval methodology for the 324-event parity lifetime analysis); (3) baseline comparisons, while appropriate, rely heavily on self-citation without independent verification of the comparison metrics; (4) absence of a systematic failure-modes discussion. The paper would benefit from explicit tolerance specifications in performance claims and a data-availability statement with recomputation scripts.

Research rigour score: 6.5/10 — The experimental methodology and results are credible and represent genuine advances, but the paper lacks the reproducibility infrastructure and explicit uncertainty quantification that strict audit standards require.

8.1.4 4. Three highest-leverage improvements

- Add data availability and reproducibility infrastructure:** Include a supplementary package with raw measurement data (time traces, conductance sweeps, TGP outputs) in a standard format (HDF5/JSON) and a script that derives every numerical claim (Z , T , EM bounds, periodicity estimates) from this data. This would transform the “Auditable claims” item from FAIL to PASS.
- Provide explicit uncertainty methodology and confidence intervals:** For the key $Z = 22 \pm 1$ s result, specify the statistical framework (maximum likelihood estimation on exponential distribution? Bayesian posterior?), report the 95% CI explicitly, and address potential systematic uncertainties (threshold selection sensitivity, drift during long traces). For the $EM < 1 \text{ } \mu\text{eV}$ claim, quantify the confidence level and detector resolution floor more precisely.
- Include a dedicated “Limitations and Failure Modes” section:** Systematically document parameter regimes where the device does not achieve the reported performance (e.g., regions outside the identified cluster in Fig. 5, conditions under which poisoning rates increase, sensitivity to magnetic field misalignment), and discuss any instances where the rf spectroscopy method failed to identify low-energy states. This honest reporting would significantly strengthen the paper’s credibility for rigorous reviewers.

8.2 majorana2 — Reviewer panel 5 voices

Model: claude-opus-4-5-20251101 · 5,072 output tokens · 121.7 s

8.2.1 Voice 1 — Reviewer 1 (Physics correctness)

This manuscript reports on an InAs–Pb hybrid device demonstrating parity lifetimes of approximately 20 seconds, which would represent a substantial improvement over the approximately 1–12 millisecond lifetimes observed in prior aluminum-based tetron devices. The physical premise is straightforward and well-motivated: replacing aluminum ($\sim 300 \text{ } \mu\text{eV}$) with lead ($\sim 1.3 \text{ meV}$) should increase both the parent superconducting gap and, consequently, the topological gap, thereby suppressing both thermal excitations and quasiparticle poisoning. The authors claim to have achieved top-quintile topological gaps of $\sim 70 \text{ } \mu\text{eV}$, compared to $\sim 30 \text{ } \mu\text{eV}$ in earlier Al-based devices. The basic physics underlying these claims is sound: the exponential dependence of Majorana hybridization on L/T , where T scales inversely with the gap in clean systems, combined with stronger electron-phonon coupling in Pb facilitating faster quasiparticle recombination, provides a coherent theoretical framework for the observed improvements.

The Hamiltonian introduced in Equation 1 describes a minimal model coupling a single low-energy wire state to a readout quantum dot. While this model is adequate for qualitative understanding, I have concerns about its application to the quantitative extraction of EM. The authors state that EM is extracted with “ μeV precision,” yet the model assumes a single isolated wire state. In the topological regime, this assumption is appropriate since 1 and 2 represent well-separated Majorana zero modes with exponentially small overlap. However, near phase boundaries or in the presence of disorder, multiple low-energy states (quasi-Majoranas) can contribute. The authors acknowledge this possibility by referencing prior work on quasi-MZMs (Ref. 61, 84), but do not explicitly demonstrate that their measurements exclude such scenarios. The claim that $\text{EM} < 1 \text{ } \mu\text{eV}$ throughout the identified topological region would be substantially strengthened by showing the parity-independence of peak width and amplitude predicted by Equation 2 in the deep topological limit where $tR \rightarrow 0$.

The induced gap measurements deserve careful scrutiny. The authors report $\sim 570 \text{ } \mu\text{eV}$ in the lowest sub-band regime at zero field for the nanowire geometry, which they claim is consistent with simulations accounting for transverse confinement. However, the proximity-induced gap measured in the 2D geometry (Fig. 2b) is $400 \text{ } \mu\text{eV}$. While transverse confinement can indeed enhance the induced gap, the 42% increase from 400 to $570 \text{ } \mu\text{eV}$ seems substantial and warrants explicit presentation of the simulation methodology. Additionally, the relationship between the parent Pb gap ($\text{Pb } 1.3 \text{ meV}$) and the induced gap is governed by interface transparency and the relative density of states in superconductor and semiconductor. The 10 nm Pb film thickness is specified, but the interface characterization that would support claims of “hard” induced gap (absence of subgap states) relies on a citation to prior Coulomb island work (Ref. 53) rather than direct measurement in the present device geometry.

The spin-orbit coupling value $\sim 12\text{--}16 \text{ meV nm}$ is extracted from Shubnikov-de Haas oscillations in Pb-proximitized nanowires according to Fig. 2c. The extraction procedure references Ref. 78, but the actual data shown appears to come from van der Pauw devices with fixed density. The authors acknowledge in their footnote (Ref. 77) that this value combines 2D measurements under Pb with weak anti-localization measurements in Hall bars without Pb. This indirect approach introduces systematic uncertainty that is not propagated into the claimed precision. Given that the topological phase boundary depends sensitively on the ratio of spin-orbit energy to Zeeman energy, more direct characterization of α in the actual device geometry would strengthen the conclusions about phase diagram extent.

Questions for Authors:

1. Can you provide explicit numerical simulations demonstrating the 42% enhancement of induced gap from transverse confinement, including the specific model assumptions about interface transparency?
2. In the topological region identified in Fig. 5, do the quantum capacitance peaks exhibit parity-independent width and amplitude as predicted by Equation 2 when $tR \rightarrow 0$, and if so, can you quantify the upper bound on tR/tR ?
3. What is the estimated systematic uncertainty in the spin-orbit coupling arising from the indirect measurement approach combining van der Pauw and Hall bar data?

Verdict: 7/10 — minor-revisions

8.2.2 Voice 2 — Reviewer 2 (Algorithmic novelty)

The central novelty claim of this work is the demonstration that increasing the superconducting gap translates directly into improved topological qubit performance, specifically through the replacement of aluminum with lead in superconductor-semiconductor hybrid devices. While this is indeed a longstanding prediction of topological quantum computing theory, the question of algorithmic or methodological novelty requires careful examination. The rf-based wire spectroscopy technique presented in Section 3 represents a genuine methodological advance over prior DC transport-based tuning protocols, enabling parallel characterization compatible with scalable architectures. This addresses a practical bottleneck identified in earlier work and represents the most novel technical contribution of the manuscript.

Comparing against recent literature, the Microsoft Quantum group’s own prior work on InAs-Al hybrid devices passing the topological gap protocol (Ref. 60, Phys. Rev. B 2023) and interferometric single-shot parity measurements (Ref. 61, Nature 2025) establishes the immediate baseline. The present work demonstrates a three-order-of-magnitude improvement in parity lifetime (20 s versus 1–12 ms), which is substantial. However, the Kanne et al. Nature Nanotechnology 2021 work (Ref. 53) already demonstrated epitaxial Pb on InAs nanowires with 2e-periodic charging patterns consistent with hard induced gap. The key advance here is not the material system per se but the integration into a multi-tetron array geometry with demonstrable parity protection. Recent work by Song et al. (Nano Lett. 2025, Ref. 55) and Zhang et al. (Nano Lett. 2026, Ref. 56) on PbTe nanowires represents related but distinct materials approaches that do not yet demonstrate comparable parity lifetimes in qubit-relevant geometries.

The claimed parity lifetime of $Z = 22 \pm 1$ s extracted from exponential fits to dwell time distributions (Fig. 6e) with $N = 324$ events raises questions about statistical power for distinguishing between competing decay models. The authors assume a homogeneous Poisson process, but non-exponential distributions could arise from time-varying quasiparticle densities or multiple competing poisoning mechanisms with different rates. With only 324 events aggregated across multiple measurements, distinguishing exponential from, say, stretched exponential or power-law tails would require careful model comparison that is not presented. The claim that “some instances reaching minute-scale” in the abstract is particularly concerning—this suggests significant variability that should be characterized systematically rather than highlighted anecdotally.

The comparison framework for establishing Pareto dominance over prior devices is implicitly presented but not rigorously constructed. The key performance metrics appear to be T (topological gap), EM (Majorana splitting), and Z (parity lifetime). For a Pareto improvement, the InAs-Pb devices should dominate on at least one metric without regression on others. The claimed improvements are T: 70 μ eV versus 30 μ eV (2.3 \times improvement), EM: < 1 μ eV (comparable or better), Z: 20 s versus 1–12 ms (>1000 \times improvement). However, the authors do not present direct measurements of X (X-parity lifetime) for the Pb-based devices, instead stating that “investigating these effects will be an important direction for future work.” This is a significant gap since X scales as EM^2 , and improved Z without corresponding improvement in X would not constitute strict Pareto dominance for all qubit operations.

Questions for Authors:

1. Have you performed formal model selection (e.g., Akaike information criterion) comparing exponential versus non-exponential dwell time distributions for the parity switching data?
2. Can you provide even preliminary X measurements for the Pb-based devices to establish whether the claimed improvements extend to X-parity operations?
3. How do the fabrication yield and reproducibility of the InAs-Pb devices compare to the Al-based baseline, given that scalability is a central motivation?

Verdict: 6/10 — major-revisions

8.2.3 Voice 3 — Reviewer 3 (Empirical evidence)

The statistical treatment of the parity lifetime measurement (Fig. 6) requires more rigorous analysis than currently presented. The authors report $Z = 22 \pm 1$ s based on exponential fitting of $N = 324$ dwell intervals aggregated across multiple measurements. The quoted uncertainty of ± 1 s appears to represent only the fit

uncertainty, not the full sampling variability. For a Poisson process with true rate λ , the maximum likelihood estimator of the mean dwell time has standard error $\sqrt{N} \approx 22/324 \approx 1.2$ s, which is consistent with the reported uncertainty. However, the aggregation across “multiple measurements” (stated as 9 additional time traces beyond the example in Fig. 6d) and across a range of $V_{wp} = 1.365 \pm 0.13$ V $\approx 10 \mu\text{V}$ introduces systematic variability that should be characterized. If the true parity lifetime varies with gate voltage even within this 20 μV range, the aggregated distribution would exhibit over-dispersion relative to a homogeneous Poisson model.

The claim of “some instances reaching minute-scale” lifetime requires quantification. Examining Fig. 6e, the histogram extends to approximately 140 s with visible counts in the 100–140 s bins. However, for $\lambda = 22$ s, the probability of observing a dwell time exceeding 100 s is $\exp(-100/22) \approx 1\%$, predicting approximately 3 events in a sample of 324. The apparent presence of multiple events at 100+ s could indicate either a heavy tail inconsistent with exponential distribution or simply sampling variability. A quantile-quantile plot against the exponential distribution would clarify whether the tail behavior is anomalous.

The rf-based wire spectroscopy protocol (Section 3, Figs. 4–5) represents a systematic approach to identifying low-energy wire states, but the thresholding procedure introduces subjective elements that should be characterized through sensitivity analysis. The signal-to-noise threshold $S/N > 2.5$ and the requirement that at least 7/10 wire cutter configurations show above-threshold signal are reasonable but arbitrary. The authors acknowledge that “the exact boundary of the identified region is in general sensitive to the details of the thresholds used,” but do not quantify this sensitivity. Varying these thresholds and reporting the resulting changes in the identified topological region would strengthen confidence that the conclusions are robust.

The manuscript lacks a systematic failure-modes section discussing devices or measurements that did not work as expected. While Section 3 mentions that “additional sign switches are interpreted to be the result of external quasiparticle poisoning events,” this is the only acknowledgment of imperfect behavior. Questions arise naturally: What fraction of devices in the multi-tetron array achieved the reported performance? Were there systematic differences between the four tetrons (AA, AB, BA, BB) in the unit cell? Did all nanowires in the device exhibit comparably low EM, or is the top wire of the BA tetron a particularly favorable case? The reproducibility claims implicit in the scalability discussion would be better supported by explicit characterization of device-to-device and wire-to-wire variability within the array.

The data provenance and reproducibility infrastructure are not described. While the authors reference specific gate voltages (e.g., $V_{wp} = 1.365 \pm 0.13$ V) and magnetic fields (e.g., $B = 3.8$ T, B_z in Fig. 5), there is no mention of data availability statements, analysis code, or raw data repositories. For a result with significant implications for quantum computing scalability, the ability to independently verify numerical claims from archived data and analysis scripts would substantially enhance credibility. Even if full data release is not feasible for proprietary reasons, describing the audit trail from raw measurements to reported numbers would address concerns about replicability.

Questions for Authors:

1. Can you provide a quantile-quantile plot comparing the observed dwell time distribution against the fitted exponential, particularly characterizing the tail behavior that gives rise to “minute-scale” instances?
2. What is the wire-to-wire and device-to-device variability in EM and Z across the multi-tetron array, and what fraction of measured wires fall within the claimed performance specifications?
3. Is there a data availability statement or analysis code repository that would enable independent verification of the reported numerical results?

Verdict: 6/10 — major-revisions

8.2.4 Voice 4 — Devil’s Advocate

This manuscript exhibits a concerning pattern of presenting aspirational claims with insufficient supporting evidence, while downplaying or omitting information that would complicate the triumphalist narrative. Let me be specific about the most serious problems that the other reviewers have been too generous in overlooking.

First, the three-order-of-magnitude improvement in parity lifetime is based on a single wire in a single device, measured over an unspecified total duration, with results aggregated across an unspecified number of measurement sessions potentially spanning days. The authors casually mention that “the measurements in

Fig. 4 and Fig. 5 were taken several days apart” with “some small shifts in gate voltages expected” (Ref. 85), revealing that device stability over relevant timescales is not actually demonstrated. If gate voltage drift of hundreds of microvolts occurs between measurement sessions, how can we be confident that the 20-second parity lifetime would be maintained over the operational timescales required for quantum computation? The entire premise of using parity lifetime as a qubit metric assumes stable operating conditions, yet stability is neither demonstrated nor even claimed.

Second, the claimed topological gap of $T \approx 70 \text{ } \mu\text{eV}$ (top quintile) is presented as a definitive improvement, but this number comes from the Topological Gap Protocol applied to test device structures (Fig. 3c), not to the actual multi-tetron array (Fig. 1) in which the parity measurements were performed. The TGP measurement was done on a “3 μm -long nanowire test structure” according to the figure caption, whereas the tetron nanowires are 3.5 μm long with different geometry including the narrow backbone junction. The authors provide no evidence that the TGP results transfer to the actual qubit geometry. This is not a minor caveat—the relationship between test structure performance and integrated device performance is the central question for any claim about scalability.

Third, the rf-based wire spectroscopy method, presented as enabling “scalable tuning,” has only been demonstrated on one wire of one tetron. The multi-tetron array shown in Fig. 1 contains eight nanowires (two per tetron, four tetrans), yet all the detailed measurements (Figs. 4–6) focus exclusively on the top wire of the BA tetron. Where is the data from the other seven wires? If this is truly a “prototype unit cell for multi-tetron devices” suitable for “scaling to larger qubit arrays,” why is there no demonstration of parallel tuning and characterization? The closest the authors come is mentioning that “all of the QDs have been designed to have plunger lever arms in the range 0.4–0.45 meV/mV,” which is a design claim, not an experimental demonstration.

Fourth, the attribution of improved parity lifetime to the Pb-Al material difference is correlational, not causal. The authors changed multiple variables simultaneously: the superconductor (Al to Pb), the substrate (implied InP to GaSb), the quantum well composition (adding InAsSb), and presumably numerous fabrication details. While they argue that the larger Pb gap and stronger electron-phonon coupling are responsible for the improvement, they provide no controlled experiment varying only the superconductor while holding other factors constant. The device measured here is not a simple material substitution—it is a complete redesign. Alternative explanations, such as reduced defect density from the GaSb substrate or reduced quasiparticle injection from improved shielding, cannot be excluded.

The manuscript’s discussion of implications for fault-tolerant quantum computing is premature to the point of being misleading. The abstract claims that “non-equilibrium quasiparticles no longer limit qubit operations in our devices,” but no actual qubit operations are demonstrated. There is no coherent manipulation, no gate fidelity measurement, no demonstration of even a single logical operation. The parity measurement is a necessary but not sufficient component of a functioning qubit. By the same logic, I could claim that a very stable piece of iron “no longer limits” magnetic memory operations because it maintains its magnetization for years—but this says nothing about whether it can function as a memory element in an actual computing system.

Recommendation: major-revisions

The paper presents genuinely interesting results on material development and measurement techniques for Majorana-based devices, but the claims about qubit performance and scalability are substantially oversold relative to the evidence presented. Major revision is required to either scale back the claims to match the evidence or to provide substantially more data supporting the scalability and reproducibility assertions.

8.2.5 Voice 5 — Editor-in-Chief synthesis

Having considered all four reviews, I find a manuscript that presents legitimate scientific advances alongside claims that substantially exceed the evidentiary support. The disagreement between reviewers reflects genuine tension in the paper between high-quality materials characterization and overreaching implications for quantum computing.

Reviewer 1 finds the physics framework sound but identifies gaps in the quantitative justification, particularly regarding induced gap simulations, spin-orbit coupling extraction methodology, and the connection between the minimal Hamiltonian model and the claim of μeV -resolution EM extraction. These are addressable through additional analysis and clearer presentation of methodology, consistent with minor revisions. Reviewer 2 raises more serious concerns about the novelty claim structure, correctly identifying that the absence of X measurements for Pb-based devices undermines the Pareto dominance argument for complete qubit performance. This

is a substantive gap that requires additional experimental data. Reviewer 3’s concerns about statistical rigor, threshold sensitivity in the tuning protocol, and absence of failure-mode documentation are well-founded and align with best practices for high-impact experimental claims.

The Devil’s Advocate raises the most damaging critique: the disconnect between the claimed “scalable multi-tetron array” architecture and the actual measurement scope, which focuses exclusively on one wire of one tetron. This criticism is valid and substantially undermines the paper’s central narrative about demonstrating a pathway to fault-tolerant quantum computing. The observation that TGP measurements come from test structures rather than the actual qubit device is particularly concerning, as it breaks the evidentiary chain linking material improvements to qubit performance. However, I do not agree with the implicit suggestion that the entire line of reasoning is invalid—the improvements in parity lifetime are real and significant, even if the scalability claims are premature.

Reconciling these perspectives, I find that the manuscript would be appropriate for PRX Quantum after major revisions. The core scientific contribution—demonstration of dramatically improved parity lifetimes in Pb-based devices combined with a scalable rf-based tuning methodology—is significant and timely. However, the presentation must be substantially revised to align claims with evidence. The scalability narrative should be tempered to acknowledge that multi-wire, multi-tetron characterization remains future work. The failure-mode and reproducibility documentation needs expansion. And the statistical treatment of parity lifetime measurements requires the rigor appropriate for a flagship result.

Must-fix items before resubmission, ordered by severity:

1. Provide characterization data from at least one additional wire in the multi-tetron array (preferably from a different tetron than BA) demonstrating comparable EM and parity lifetime, or revise all claims about scalability and array-level performance to explicitly acknowledge single-wire scope.
2. Include either preliminary X measurements for the Pb-based devices or remove all claims about implications for X-parity operations and general qubit performance, explicitly scoping the results to Z-parity measurements only.
3. Present quantitative model comparison (AIC or similar) between exponential and alternative dwell time distributions, including explicit treatment of the “minute-scale instances” tail behavior and characterization of measurement-to-measurement variability in extracted Z.
4. Perform and report sensitivity analysis on the rf tuning protocol thresholds, quantifying how the identified topological region changes with S/ threshold and cutter configuration requirements.
5. Either present TGP measurements on the actual tetron device in which parity measurements were performed, or clearly separate test structure characterization from device-level performance with explicit discussion of the expected transfer of properties.
6. Add a data availability statement specifying the accessibility of raw measurement data and analysis code, consistent with PRX Quantum policies on reproducibility.

8.2.6 Vote table

Voice	Recommendation	Confidence 1-10
Reviewer 1	minor-revisions	7
Reviewer 2	major-revisions	6
Reviewer 3	major-revisions	6
Devil’s Advocate	major-revisions	8
Editor-in-Chief	major-revisions	7

8.3 majorana2 — Logical-fallacy report

Model: claude-opus-4-5-20251101 · 3,112 output tokens · 62.7s

Fallacy: cherry-picked-baseline **Severity:** high **Location:** Section 1 (Introduction), paragraph 2 and throughout **Evidence:** “In earlier Al-based devices, we observed typical top quintile gaps of $T \approx 30 \text{ } \mu\text{eV}$ [59–61]. In contrast, in the InAs–Pb devices studied here we observe a top quintile gap of $T \approx 70 \text{ } \mu\text{eV}$.” **Why it’s the fallacy:** The manuscript exclusively compares its Pb-based devices against the authors’ own earlier Al-based devices from their own lab (refs 59-61). While refs 53-56 mention other groups’ Pb-based nanowire work, the paper dismisses this prior art with the vague claim that they have “gone beyond previous work which has incorporated larger gap superconductors into nanowire devices [53–56]” without providing any quantitative comparison of topological gaps, parity lifetimes, or other metrics against these published baselines. This selective comparison against only weaker internal baselines while ignoring potentially stronger external results constitutes cherry-picking. **Suggested fix:** Add a table comparing quantitative metrics (T , parity lifetime, EM) against the specific results from refs 53-56 and any other relevant Pb-based or high-gap superconductor nanowire studies. If those studies didn’t report comparable metrics, state this explicitly.

Fallacy: hasty-generalization **Severity:** medium **Location:** Section 4 (Interferometric Parity Readout), paragraph on parity lifetime **Evidence:** “To quantify the lifetime we classify the data by fitting a Gaussian mixture model... By aggregating multiple measurements, we observe a total of $N = 324$ dwell intervals. The data are consistent with a single exponential distribution... We extract a characteristic parity lifetime and corresponding fit uncertainty of $Z = 22 \pm 1 \text{ s}$ ” **Why it’s the fallacy:** The 20+ second parity lifetime claim—which is the paper’s headline result—is derived from measurements on a single nanowire of a single tetron in a single device. The abstract and conclusions generalize this to “InAs–Pb tetron devices” broadly, but there is no demonstration that this result is reproducible across multiple devices, multiple tetrans, or even multiple wires within the same device. The measurement also appears to be taken at a specific optimized operating point ($V_{wp} = 1.365 \text{ } 13 \text{ V}$). **Suggested fix:** Either present parity lifetime statistics from multiple devices/tetrans/wires, or qualify claims to state “In a single nanowire of one tetron, we measured...” and explicitly note this as a demonstration requiring broader validation.

Fallacy: conflated-regimes **Severity:** medium **Location:** Section 5 (Discussion and Outlook), final paragraphs **Evidence:** “The multi-tetron array presented here functions as a modular ‘unit cell’ for a larger architecture; it can be tiled into much larger qubit arrays (e.g., a 12-qubit array) without altering the underlying control or readout approach. The strong parity protection observed in our tetron prototype suggests that, even as the system scales up, each qubit will remain well isolated from non-equilibrium quasiparticles” **Why it’s the fallacy:** The paper extrapolates from a single tetron measurement (one wire, one parity lifetime measurement) to claims about how a 12-qubit array would behave. This conflates the regime of a single isolated device with the regime of scaled arrays where crosstalk, thermal load from additional control lines, and other scaling-specific noise sources may dominate. No evidence is provided that scaling preserves these properties. **Suggested fix:** Reframe as: “Demonstrating similar performance in scaled arrays remains an important open challenge. Potential scaling concerns include [list specific concerns] which will require investigation in multi-qubit devices.”

Fallacy: active-space-handwave **Severity:** medium **Location:** Section 5 (Discussion and Outlook) **Evidence:** “Looking ahead, the values of EM achieved here are expected to have beneficial implications for Pauli-X measurements, whose characteristic switching time X scales as $EM^{\dot{}}$. Deep in the topological regime, $EM \propto T \exp(L/T)$... Increasing the NW length L exponentially suppresses EM and thus dramatically extends X , which suggests that X could be more than an order of magnitude longer than in previous devices.” **Why it’s the fallacy:** The paper claims that X “could be more than an order of magnitude longer” based on theoretical scaling arguments, but does not actually perform X -parity measurements or demonstrate this. The claim of generalizing to X measurements is made without running the experiment, which constitutes an active-space handwave—claiming a result extends to a regime not actually tested. **Suggested fix:** Remove the speculative quantitative claim (“more than an order of magnitude”) or qualify it as: “Based on theoretical scaling relations, we predict X improvements, but experimental validation of X -parity lifetimes in Pb-based devices remains for future work.”

Fallacy: ad-hoc-precision-floor **Severity:** medium **Location:** Section 3 (RF-Based Wire Spectroscopy), discussion of EM resolution **Evidence:** “we expect the resolution of this analysis to be limited to 1 μeV ... Notably, the resolution of 1 μeV significantly exceeds that of conductance measurements which can resolve EM of about the half-width-half-max of a temperature broadened conductance peak EM $1.76k_B T$ 7.6 μeV for $T = 50$ mK.” **Why it’s the fallacy:** The paper claims EM values “below our 1 μeV resolution” as evidence of topological behavior, but this resolution floor is derived from DAC resolution and lever arm estimates, not from demonstrated noise floors or calibration standards. The comparison to “7.6 μeV ” for conductance measurements also involves mixing different measurement modalities and conditions without rigorous cross-calibration. Claiming sub-resolution EM values as a positive result approaches the ad-hoc-precision-floor fallacy. **Suggested fix:** Provide explicit calibration of the 1 μeV resolution claim with known energy scales, or reframe as: “We observe EM consistent with zero within our measurement resolution of approximately 1 μeV , though we cannot exclude finite EM below this threshold.”

Fallacy: appeal-to-authority **Severity:** medium **Location:** Section 1 (Introduction), opening paragraph **Evidence:** “Recently, we presented a roadmap [1] to fault-tolerant quantum computation using topological qubits [2–4] built around Majorana zero modes (MZMs) in superconductor-semiconductor hybrid devices [5–9]. Our roadmap draws on concepts explored in Refs. 10–39.” **Why it’s the fallacy:** The paper opens by citing 39 references in the first paragraph, with 30 of them (refs 10-39) bundled into a single “draws on concepts” citation. This mass citation serves more as an appeal to the authority and weight of the cited literature than as specific technical justification. The roadmap ref [1] is the authors’ own prior publication, creating a self-referential authority structure. **Suggested fix:** Cite specific concepts from specific papers where they are used technically, rather than bundling 30 references as general background authority. If a roadmap citation is needed, briefly state which specific elements are being implemented.

Fallacy: equivocation **Severity:** medium **Location:** Throughout, particularly Abstract and Section 4 **Evidence:** Abstract: “parity lifetime of 20 s with some instances reaching minute-scale”; Section 4: “intrinsic Z-parity lifetime of the device which far exceeds the measurement time” **Why it’s the fallacy:** The paper uses “parity lifetime” to refer to multiple related but distinct quantities: (1) Z-parity lifetime in an interference measurement, (2) quasiparticle poisoning time, and (3) intrinsic parity switching time. The headline “20 second parity lifetime” refers specifically to Z-parity in one measurement configuration, but the discussion sometimes conflates this with general qubit lifetime or poisoning immunity. The phrase “some instances reaching minute-scale” is particularly vague about what operational definition of parity lifetime is being used. **Suggested fix:** Define “Z-parity lifetime” precisely at first use and use consistent terminology throughout. Clarify whether “minute-scale” instances represent statistical fluctuations in the exponential distribution or a different measurement condition.

8.4 majorana2 — Argument-structure audit

Model: claude-haiku-4-5-20251001 · 3,083 output tokens · 73.5 s

ARGUMENT STRUCTURE REPORT

8.4.1 Executive summary

Overall verdict: PUBLISH-READY **Claim–proof gap:** NONE — The paper claims that replacing Al with Pb in tetron devices yields larger topological gaps and longer parity lifetimes, and demonstrates exactly this through material characterization, TGP analysis, and time-resolved parity measurements. **CME balance:** BALANCED — C:25% / M:30% / E:45% **Narrative debts:** 5 total (0 load-bearing) **Sequencing:** EVIDENTIAL

8.4.2 Argument map

P1 (premise, §1): Topological protection gives error rates exponentially suppressed as the topological gap increases.

P2 (premise, §1): The parent superconducting gap governs the topological gap in superconductor-semiconductor hybrids [Refs. 4052].

P3 (premise, §2): Pb has a parent gap $\Delta_{\text{Pb}} = 1.3 \text{ meV}$ vs Al $\Delta_{\text{Al}} = 300 \text{ } \mu\text{eV}$.

P4 (premise, §2): The GaSb substrate enables lattice-matched growth and larger spin-orbit coupling ($\sim 12 \text{ meV}$).

I1 (intermediate, from P2+P3+P4): InAsPb devices should exhibit larger topological gaps than Al-based devices.

E1 (evidence, §23): TGP analysis shows top-quintile $T \sim 70 \text{ } \mu\text{eV}$ in Pb devices vs $\sim 30 \text{ } \mu\text{eV}$ in Al devices.

E2 (evidence, §3): Localization lengths exceed $1 \text{ } \mu\text{m}$, confirming low disorder.

E3 (evidence, §3): Zero-bias peaks persist over $>0.5 \text{ T}$ field range at both wire ends.

I2 (intermediate, from I1+E1+E2+E3): The Pb-based platform achieves a robust topological phase with enlarged topological gaps.

P5 (premise, §1): Non-equilibrium quasiparticles limit parity lifetime independently of T .

P6 (premise, §4): Pb's higher gap suppresses Cooper pair breaking and enhances recombination.

E4 (evidence, §4): Measured Z-parity lifetime $Z = 22 \pm 1 \text{ s}$ via interferometric readout with $h/2e$ periodicity.

E5 (evidence, §4): Prior Al-based devices showed $Z \sim 112 \text{ ms}$ [Refs. 59, 61].

I3 (intermediate, from P5+P6+E4+E5): Parity lifetime improved by >3 orders of magnitude due to Pb substitution.

E6 (evidence, §3): rf spectroscopy shows $EM < 1 \text{ } \mu\text{eV}$ across extended parameter regimes.

I4 (intermediate, from E6): Majorana hybridization is suppressed below measurement resolution.

C (conclusion, from I2+I3+I4): Increasing the excitation gap via Pb substitution directly translates to improved qubit performance.

Unsupported leaps: None identified. Each intermediate claim follows from stated premises plus presented evidence.

Unstated premises: 1. The parity lifetime measurement faithfully reflects quasiparticle poisoning rate rather than alternative relaxation mechanisms (partially addressed in §4 discussion of Poisson statistics). 2. The single measured tetron is representative of the array; no cross-device statistics are shown for parity lifetime. 3. Floating tetrons will have comparable or better parity lifetimes (acknowledged as expectation, not demonstrated).

8.4.3 Section A: Controlling idea

(a) STATED CLAIM

“Here, we experimentally validate this principle in an InAs–Pb tetron device via interferometric single-shot parity measurements. By replacing aluminum with the higher-gap superconductor lead in our superconductor-semiconductor hybrid devices, we have improved the robustness of our topological phase.”

The strongest single sentence appears in §5:

“Our results confirm a central premise of topological quantum computing: increasing the excitation gap dramatically reduces error mechanisms and improves qubit performance.”

(b) DEMONSTRATED CONCLUSION

An InAs–Pb tetron device exhibits a measured topological gap of ~ 70 μeV (vs ~ 30 μeV in Al), Majorana splitting EM < 1 μeV , and Z-parity lifetime of ~ 22 s (vs ~ 1 – 12 ms in Al), directly demonstrating that substituting a higher-gap superconductor yields quantitative performance improvements across multiple metrics.

(c) CLAIM–PROOF GAP

NONE — The abstract and introduction claim that increasing the gap improves device performance, and the evidence architecture establishes exactly this through: (i) material characterization showing higher induced gap, (ii) TGP analysis showing larger topological gap, (iii) rf spectroscopy showing low EM, and (iv) time-resolved parity measurement showing ~ 22 s lifetime. The comparison to prior Al-based devices (Refs. 59, 61) anchors the improvement claim. The paper does not overclaim universality or fault-tolerance; it states these are “implications” and “expectations” for future work.

8.4.4 Section B: CME proportionality**CLAIM (C): STRONG — $\sim 25\%$**

Claims are made in §1 (introduction) and §5 (discussion): improved topological gap, suppressed EM, orders-of-magnitude improvement in parity lifetime, scalability implications. Claims are appropriately scoped to what is demonstrated and clearly distinguish demonstrated results from extrapolated expectations (e.g., “we expect that parity lifetimes in floating tetrons could be even longer”).

MECHANISM (M): STRONG — $\sim 30\%$

§1 provides the theoretical framework (topological protection scales with gap, quasiparticle poisoning independent of T). §2 explains why Pb + GaSb substrate should yield improvements (higher parent gap, lattice-matched growth, enhanced spin-orbit). §3 (rf spectroscopy) derives the single-state model Hamiltonian (Eq. 1) and quantum capacitance response (Eq. 2). §4 explains the physical origin of parity lifetime enhancement (harder to break Cooper pairs in Pb, faster recombination). The mechanism story is complete and quantitative.

EVIDENCE (E): STRONG — $\sim 45\%$

The paper is evidence-rich: - Material characterization: Fig. 2 (induced gap 400–570 μeV , mobility $> 350,000$ cm^2/Vs , spin-orbit 12 ± 2 meV nm) - TGP analysis: Fig. 3 (T ~ 70 μeV , topological region > 1.1 $\text{mV}\mu\text{T}$, zero-bias peaks over > 0.5 T) - rf spectroscopy: Figs. 4–5 (EM extraction with ~ 1 μeV resolution, correlation maps) - Parity measurement: Fig. 6 ($h/2e$ periodicity, $Z = 22 \pm 1$ s from exponential fit to $N=324$ dwell intervals)

Verdict: BALANCED

C:25% / M:30% / E:45%. No dimension dominates by more than ~ 20 points; evidence is the largest component but appropriately so for an experimental paper.

8.4.5 Section C: Narrative-debt register

Promise type	Promise (location)	Status	Load-bearing?
EVIDENCE	“we experimentally validate this principle” (abstract)	FULFILLED	N/A
PROMISE	“we have developed an rf measurement technique that resolves low-energy wire-end states and directly measures their energy splitting with μeV precision” (abstract)	FULFILLED — §3 demonstrates resolution ~ 1 μeV	No
EVIDENCE	“We employ this technique to bring up a device in a multi-tetron array” (abstract)	FULFILLED — §3–4 operate on the BA tetron in the array	No
PROMISE	“Further time-resolved measurements reveal a characteristic parity switching time of ~ 20 s with some instances reaching minute-scale” (abstract)	PARTIAL — $Z = 22 \pm 1$ s is shown; “minute-scale” instances are not explicitly displayed in Fig. 6(e) distribution	Cosmetic

Promise type	Promise (location)	Status	Load-bearing?
RHETORICAL QUESTION	“we discuss potential implications for the fidelity of Pauli measurements” (abstract)	PARTIAL — §5 discusses qualitatively but does not quantify fidelity	Cosmetic

Total narrative debts: 5 (0 load-bearing)

The “minute-scale” claim is supported by the tail of the exponential distribution but no individual trace is highlighted; this is minor. The fidelity discussion in §5 is qualitative, but the abstract uses “discuss” rather than “quantify,” so no strong expectation is violated.

8.4.6 Section D: Sequencing diagnosis

Verdict: EVIDENTIAL

The paper follows a classic escalation structure: 1. §1 (Introduction): Establishes premises (topological protection, roadmap, error mechanisms) 2. §2 (Device design and materials): Describes platform and justifies why it should work 3. §3 (rf spectroscopy): Introduces new characterization method, demonstrates low EM 4. §4 (Parity readout): Presents the headline result ($Z \sim 20$ s) 5. §5 (Discussion): Synthesizes and extrapolates

The strongest result (parity lifetime) appears late, after the evidential scaffolding is complete. The abstract front-loads the result, which is standard for physics journals and does not constitute headline-first body organization.

Proposed resequencing moves: None required. The current order supports evidential escalation effectively. One optional refinement:

Current position	Optimal position	Benefit
§5 paragraph on X scaling (“the values of EM achieved here are expected to have beneficial implications...”)	Could appear earlier in §4 as a framing paragraph	Would connect EM measurements (§3) to parity measurements (§4) more tightly before presenting data

This is minor and not essential.

8.4.7 Section E: Structural gaps

Missing analysis	Where it should appear	CME dimension strengthened
Cross-device statistics for parity lifetime	§4 or supplementary	Evidence — currently $N=1$ tetron; showing variance across array would strengthen scalability claim
Error bars on T values	§3 (Fig. 3c)	Evidence — “top quintile” is stated but distribution not shown
Direct comparison table: Al vs Pb metrics	§5 or §2	Claim — would make the “3 orders of magnitude” comparison more immediate
Noise model for CQ telegraph noise	§4	Mechanism — would formalize the Poisson assumption and bound systematic errors
X measurement or projection	§5	Evidence — discussed qualitatively but not measured; would discharge the Pauli-X implication promise

None of these gaps are critical; the paper is complete for its stated scope.

8.4.8 Summary diagnosis

The paper claims that substituting lead for aluminum in a tetron device yields dramatically improved topological protection, and demonstrates exactly this: a measured topological gap $\sim 2\epsilon$ larger, Majorana splitting below 1 μeV , and parity lifetime $>1000\epsilon$ longer than prior Al-based devices. The argument is evidentially complete, with strong material characterization (TEM, Hall bars, SdH), protocol development (rf spectroscopy), and direct measurement ($h/2e$ -periodic parity readout with $Z = 22\text{ s}$). The CME balance is appropriate for an experimental paper; no dimension is absent or critically thin. Narrative debts are cosmetic. Sequencing is evidential. The paper is what it claims to be: an experimental validation that increasing the superconducting gap improves topological qubit performance. No restructuring is required; minor additions (cross-device statistics, explicit comparison table) would strengthen but are not essential.

8.5 majorana2 — Numeric-claim registry (deterministic)

Numeric-Claim Registry Audit Paper: `majorana2.pdf` — 2 sections indexed, 0 distinct ground-truth values (from none).

Not applicable: no Results/Experiments/Tables section was identified, so there is no in-paper ground truth to gate against. This is expected for letter-style papers without named section headings or for lossy PDF extraction. Run against the LaTeX/markdown source, or pass `-strict-on` with this paper’s actual section names.

Clean. Every numeric in the non-strict sections traces back to a value present in Results / Experiments / Tables.

8.6 majorana2 — Disclosure audit

Model: `claude-opus-4-5-20251101` · 1,781 output tokens · 33.7s

DISCLOSURE AUDIT REPORT

8.6.1 Compliance status by category

Code	Category	Status	Evidence (section or “absent”)	Required for venue?
A1	Funding statement	ABSENT	absent	Yes
A2	Competing interests / COI	ABSENT	absent	Yes
A3	Author contributions	ABSENT	absent	Yes
A4	Data availability	ABSENT	absent	Yes
A5	Code availability	ABSENT	absent	Yes
A6	Ethics / IRB approval	NOT- APPLICABLE	Pure device/experimental physics paper with no human/animal subjects	No
A7	Preprint / prior-publication status	PRESENT- COMPLETE	“arXiv:2606.03884v1 [cond-mat.mes-hall] 2 Jun 2026” on page 1	Yes
A8	Materials availability	ABSENT	absent	Yes
B1	AI-assisted text drafting	ABSENT	absent	Yes (if applicable)
B2	AI-generated images/figures	ABSENT	absent	Yes (if applicable)
B3	AI-assisted data analysis or coding	ABSENT	absent	Yes (if applicable)
B4	AI-assisted pre-submission review	ABSENT	absent	Yes (if applicable)
C1	Prior-publication warranty	ABSENT	absent	Yes

Code	Category	Status	Evidence (section or “absent”)	Required for venue?
C2	Government / institutional / export-control approval	ABSENT	absent	Yes (quantum hardware work)
C3	Third-party rights	ABSENT	absent	Yes (if applicable)
C4	Free-online-version conflict	PRESENT-INCOMPLETE	arXiv preprint exists; PRX Quantum permits arXiv but manuscript lacks explicit statement of compliance	Yes

8.6.2 Missing disclosures and warranty gaps

Submission-blocking

- **A1 (Funding statement)**: No funding sources or grant numbers disclosed. Microsoft Quantum authorship implies corporate funding but explicit statement required.
- **A2 (Competing interests)**: No COI declaration. All authors affiliated with Microsoft Quantum, which has commercial interest in quantum computing; financial/employment conflicts must be declared.
- **A3 (Author contributions)**: Group authorship “Microsoft Quantum” with footnote listing 150+ contributors but no CRediT-style role assignments.
- **A4 (Data availability)**: No statement regarding experimental data, processed datasets, or access procedures.
- **A5 (Code availability)**: References to simulations and analysis code (e.g., TGP protocol, fitting procedures) but no repository, license, or availability statement.
- **C1 (Prior-publication warranty)**: No explicit statement that manuscript is not under consideration elsewhere.

Revise-before-acceptance

- **A8 (Materials availability)**: Device fabrication involves proprietary InAs–Pb heterostructures, GaSb substrates, and specific epitaxial growth. No MTA information or availability statement provided.
- **C2 (Government/institutional/export-control approval)**: Quantum computing hardware may be subject to export controls; no pre-publication clearance statement.
- **B1–B4 (AI disclosures)**: PRX Quantum increasingly requires disclosure of AI assistance. Authors must confirm whether AI tools were used and provide explicit negative statement if not.

Editorial-discretion

- **C3 (Third-party rights)**: All figures appear original. Confirm no reuse requiring permission.
- **C4 (Free-online-version conflict)**: arXiv posting compatible with PRX Quantum policy but explicit acknowledgment recommended.

8.6.3 Submission-ready checklist

Code	Status	Fix Required
A1	FAIL	Add funding statement with grant numbers or “This work was funded internally by Microsoft Corporation with no external grant support.”
A2	FAIL	Add competing interests declaration: all authors are employees of Microsoft Corporation which is developing commercial quantum computing technology.
A3	FAIL	Add CRediT-style author contributions for the 150+ listed contributors or designate writing committee with specific roles.
A4	FAIL	Add data availability statement specifying repository/access or “Data available from corresponding author upon reasonable request.”

Code	Status	Fix Required
A5	FAIL	Add code availability statement with repository URL and license, or explicit “Code available upon reasonable request” with justification.
A6	PASS	Not applicable for device physics paper.
A7	PASS	arXiv ID present: 2606.03884v1.
A8	FAIL	Add materials availability statement addressing InAs–Pb heterostructure availability, fabrication protocols, and MTA requirements.
B1	FAIL	Add explicit statement: “AI tools were [not] used for text drafting” with model/scope if applicable.
B2	FAIL	Add explicit statement: “All figures were created without AI assistance” or disclose tools used.
B3	FAIL	Add explicit statement regarding AI use in data analysis/coding or confirm none used.
B4	FAIL	Add explicit statement regarding AI-assisted review or confirm none used.
C1	FAIL	Add warranty: “This manuscript has not been published previously and is not under consideration at another journal.”
C2	FAIL	Add statement confirming institutional/export-control pre-publication approval for quantum hardware disclosure.
C3	PASS	Requires author confirmation that all figures are original; no obvious third-party content identified.
C4	PASS	arXiv preprint policy compatible with PRX Quantum; recommend adding explicit statement.

8.7 majorana2 — Anchored revision plan

Model: claude-haiku-4-5-20251001 · 2,676 output tokens · 56.1 s

Anchored Revision Roadmap

8.7.1 1. Provide characterization data from at least one additional wire in the multi-tetron array (preferably from a different tetron than BA) demonstrating comparable EM and parity lifetime, or revise all claims about scalability and array-level performance to explicitly acknowledge single-wire scope.

Severity: 5 **Effort:** multi-day **Judges:** Reviewer 2, Reviewer 3, Devil’s Advocate, Editor-in-Chief

Source paragraph(s): ¶003, ¶008, ¶024

Quoted problem prose (verbatim from the manuscript, ≤2 sentences): > “We employ this technique to bring up a device in a multi-tetron array and perform parity measurements of one of the tetron’s hybrid nanowires (NWs).”

Judge evidence (one bullet per judge that diagnosed this; quote the judge verbatim, ≤1 sentence each):
 - Reviewer 3: “What is the wire-to-wire and device-to-device variability in EM and Z across the multi-tetron array, and what fraction of measured wires fall within the claimed performance specifications?” - Devil’s Advocate: “The multi-tetron array shown in Fig. 1 contains eight nanowires (two per tetron, four tetrans), yet all the detailed measurements (Figs. 4–6) focus exclusively on the top wire of the BA tetron.” - Editor-in-Chief: “Provide characterization data from at least one additional wire in the multi-tetron array (preferably from a different tetron than BA) demonstrating comparable EM and parity lifetime, or revise all claims about scalability and array-level performance to explicitly acknowledge single-wire scope.”

Proposed edit: Either add a supplementary figure presenting EM extraction and parity lifetime data from at least one additional wire (e.g., the bottom wire of BA or any wire from AA, AB, or BB tetron), or revise ¶003 and ¶024 to read: “We employ this technique to bring up a device in a multi-tetron array and perform parity measurements on *one nanowire of one tetron*; characterization of array-wide performance remains for future work.”

Why this works: Explicitly scoping the measurement to a single wire eliminates the mismatch between the claimed “scalable multi-tetron array” framing and the single-wire evidence actually presented.

8.7.2 2. Include either preliminary X measurements for the Pb-based devices or remove all claims about implications for X-parity operations and general qubit performance, explicitly scoping the results to Z-parity measurements only.

Severity: 4 **Effort:** 2 h (if removing claims) or multi-day (if adding measurements) **Judges:** Reviewer 2, Editor-in-Chief

Source paragraph(s): ũ024

Quoted problem prose (verbatim from the manuscript, ≤ 2 sentences): > “Looking ahead, the values of EM achieved here are expected to have beneficial implications for Pauli-X measurements, whose characteristic switching time X scales as EMš. . . . Investigating these effects will be an important direction for future work.”

Judge evidence (one bullet per judge that diagnosed this; quote the judge verbatim, ≤ 1 sentence each): - Reviewer 2: “Can you provide even preliminary X measurements for the Pb-based devices to establish whether the claimed improvements extend to X-parity operations?” - Editor-in-Chief: “Include either preliminary X measurements for the Pb-based devices or remove all claims about implications for X-parity operations and general qubit performance, explicitly scoping the results to Z-parity measurements only.”

Proposed edit: If X data are unavailable, revise ũ024 to delete the sentence “which suggests that X could be more than an order of magnitude longer than in previous devices” and replace with: “Experimental characterization of X in Pb-based devices remains an open direction; the results presented here are limited to Z-parity measurements.”

Why this works: Removes speculative quantitative claims about unmeasured quantities, aligning claims with the presented evidence.

8.7.3 3. Present quantitative model comparison (AIC or similar) between exponential and alternative dwell time distributions, including explicit treatment of the “minute-scale instances” tail behavior and characterization of measurement-to-measurement variability in extracted Z.

Severity: 4 **Effort:** 2 h **Judges:** Reviewer 2, Reviewer 3, Editor-in-Chief

Source paragraph(s): ũ003, ũ022

Quoted problem prose (verbatim from the manuscript, ≤ 2 sentences): > “Further time-resolved measurements reveal a characteristic parity switching time of 20 s with some instances reaching minute-scale.”

Judge evidence (one bullet per judge that diagnosed this; quote the judge verbatim, ≤ 1 sentence each): - Reviewer 2: “Have you performed formal model selection (e.g., Akaike information criterion) comparing exponential versus non-exponential dwell time distributions for the parity switching data?” - Reviewer 3: “Can you provide a quantile-quantile plot comparing the observed dwell time distribution against the fitted exponential, particularly characterizing the tail behavior that gives rise to ‘minute-scale’ instances?” - Editor-in-Chief: “Present quantitative model comparison (AIC or similar) between exponential and alternative dwell time distributions, including explicit treatment of the ‘minute-scale instances’ tail behavior and characterization of measurement-to-measurement variability in extracted Z.”

Proposed edit: Add to ũ022 or a supplementary section: (1) a Q-Q plot of observed dwell times vs. the exponential(=22 s) quantiles, (2) AIC values comparing exponential, stretched-exponential, and power-law fits, and (3) a table showing Z extracted from each of the 9 individual time traces with standard deviation to characterize run-to-run variability.

Why this works: Demonstrates that the exponential assumption is justified (or identifies systematic deviations), and quantifies the “minute-scale” tail relative to statistical expectation.

8.7.4 4. Perform and report sensitivity analysis on the rf tuning protocol thresholds, quantifying how the identified topological region changes with S/ threshold and cutter configuration requirements.

Severity: 3 **Effort:** 2 h **Judges:** Reviewer 3, Editor-in-Chief

Source paragraph(s): ũ020

Quoted problem prose (verbatim from the manuscript, ≤ 2 sentences): > “To identify regions associated with stable low-energy states, we evaluate, for each point in Vwp and Bz, the fraction of wire-cutter configurations for which the rf response exceeds a signal-to-noise ratio $S/ > 2.5$.”

Judge evidence (one bullet per judge that diagnosed this; quote the judge verbatim, ≤ 1 sentence each):
 - Reviewer 3: “The thresholding procedure introduces subjective elements that should be characterized through sensitivity analysis.” - Editor-in-Chief: “Perform and report sensitivity analysis on the rf tuning protocol thresholds, quantifying how the identified topological region changes with $S/$ threshold and cutter configuration requirements.”

Proposed edit: Add a supplementary figure showing the identified region of interest (analogous to Fig. 5c) for $S/$ thresholds of 2.0, 2.5, and 3.0, and for requiring 5/10, 7/10, and 9/10 cutter configurations. Include a brief statement in ũ020: “Varying the $S/$ threshold from 2.0 to 3.0 shifts the region boundary by approximately X mV in Vwp; qualitative features are preserved.”

Why this works: Demonstrates robustness (or quantifies sensitivity) of the tuning protocol to threshold choices, addressing concerns about subjectivity.

8.7.5 5. Either present TGP measurements on the actual tetron device in which parity measurements were performed, or clearly separate test structure characterization from device-level performance with explicit discussion of the expected transfer of properties.

Severity: 4 **Effort:** 30 min (if clarifying scope) or multi-day (if adding TGP data) **Judges:** Devil’s Advocate, Editor-in-Chief

Source paragraph(s): ũ012

Quoted problem prose (verbatim from the manuscript, ≤ 2 sentences): > “A representative phase diagram obtained from a NW test device is shown in Fig. 3(c), which was obtained by applying the Topological Gap Protocol (TGP) [60].”

Judge evidence (one bullet per judge that diagnosed this; quote the judge verbatim, ≤ 1 sentence each):
 - Devil’s Advocate: “The TGP measurement was done on a ‘3 μm -long nanowire test structure’ according to the figure caption, whereas the tetron nanowires are 3.5 μm long with different geometry including the narrow backbone junction.” - Editor-in-Chief: “Either present TGP measurements on the actual tetron device in which parity measurements were performed, or clearly separate test structure characterization from device-level performance with explicit discussion of the expected transfer of properties.”

Proposed edit: Add to ũ012 or the Fig. 3 caption: “The TGP data shown here were obtained from a standalone 3 μm test nanowire; TGP measurements on the integrated tetron device were not performed due to [reason, e.g., lack of transport terminals]. We expect qualitatively similar behavior in the 3.5 μm tetron wires based on [shared material stack / comparable localization lengths], though quantitative T values may differ.”

Why this works: Breaks the implicit claim that Fig. 3 directly characterizes the device in Figs. 4–6, providing transparency about the evidentiary gap.

8.7.6 6. Add a data availability statement specifying the accessibility of raw measurement data and analysis code, consistent with PRX Quantum policies on reproducibility.

Severity: 3 **Effort:** 5 min **Judges:** Reviewer 3, Editor-in-Chief

Source paragraph(s): ũ? (structural finding—no data availability section exists)

Quoted problem prose (verbatim from the manuscript, ≤ 2 sentences): > (no verbatim anchor — structural finding)

Judge evidence (one bullet per judge that diagnosed this; quote the judge verbatim, ≤ 1 sentence each):
 - Reviewer 3: “Is there a data availability statement or analysis code repository that would enable independent verification of the reported numerical results?” - Editor-in-Chief: “Add a data availability statement specifying the accessibility of raw measurement data and analysis code, consistent with PRX Quantum policies on reproducibility.”

Proposed edit: Add a new section after Acknowledgments titled “Data Availability” with text such as: “Raw measurement data and analysis scripts used to generate the figures in this work are available from the corresponding author upon reasonable request. [Alternatively: are deposited at DOI:xxxx with CC-BY license.]”

Why this works: Satisfies PRX Quantum reproducibility requirements and addresses reviewer concerns about replicability.

8.8 majorana2 — Stage-6 CQE narrative

Model: claude-haiku-4-5-20251001 · 1,089 output tokens · 28.3s

Process Summary: QuantumNovelty Run Evaluation

8.8.1 Composite Verdict

The geometric mean composite score of **23 out of 100** places this run firmly in the “**Insufficient**” tier of collaboration quality. On the standard interpretation scale, scores below 30 indicate that fundamental workflow stages were either skipped or produced incomplete outputs, rendering the run unsuitable for any downstream scientific claims. This is not a marginal miss—a score of 23 reflects systemic gaps across nearly every dimension of the evaluation framework.

To be direct: this run did not produce a valid scientific artifact. The geometric mean methodology is intentionally punishing when dimensions cluster in the low range, because poor performance in any single area can invalidate an entire research effort. Here, no dimension exceeded 40, and multiple dimensions scored below 30, indicating that the collaboration failed to establish even baseline credibility checkpoints.

8.8.2 Strongest Dimension: Communication (Score: 40)

The **Communication** dimension achieved the highest score at 40, though this must be interpreted with significant caveats. Both constituent probes—“logical fallacies absent” and “reviewer panel verdict”—scored 40, but the evidence reveals why: neither validation step actually executed. The `logical_fallacies` skill was not run, and no `review_panel.md` file was generated.

This creates an evaluation artifact where the absence of detected problems is conflated with the absence of problems themselves. A score of 40 in Communication does not indicate strong scientific prose or clear argumentation—it indicates that the tools designed to surface communication failures were never invoked. The run cannot claim communication strength; it can only claim that communication quality remains untested.

If this is the strongest dimension, it speaks to the run’s fundamental incompleteness rather than any particular competence. Communication typically improves as artifacts accumulate and undergo revision cycles. Without a paper draft (`paper.tex` exists: False), there was nothing substantive to evaluate for logical coherence or reviewer readiness.

8.8.3 Weakest Dimension: Novelty Rigour (Score: 8)

The **Novelty Rigour** dimension scored a critically low 8, driven by near-total failure in the baseline comparison infrastructure. The probe “augmented baseline catalog present” shows the `baseline_catalog` has **0 rows**—meaning no prior work was indexed against which to evaluate novelty claims. The “strict-domination comparator run” probe scored 5 because `novelty_verdict.json` was not found, indicating the comparator stage never executed or never completed.

This failure almost certainly originated in the **Literature Grounding stage** or its immediate successors. A QuantumNovelty run must first establish what constitutes known work before claiming any result is new. With an empty baseline catalog, no downstream novelty claim can survive scrutiny. Even if experimental results were generated (which other probes suggest they were not), they would be scientifically meaningless without situating them against existing methods.

The low Novelty Rigour score is particularly damaging because it undermines the entire purpose of the workflow. Quantum computing research is dense with incremental advances; claiming novelty without rigorous baseline comparison is not merely incomplete—it risks embarrassment upon peer review.

8.8.4 Three Highest-Leverage Improvements

1. Populate the Baseline Catalog Before Any Experimentation The single most impactful fix is ensuring the `baseline_catalog` contains relevant prior work before experiments begin. This is a gating dependency. The next run should include an explicit checkpoint that halts progression if `baseline_catalog` rows equal zero. Recommended implementation: integrate a mandatory literature sweep using established databases (arXiv, PubChem, prior QuantumNovelty archives) with human verification of at least 10-15 relevant baseline entries.

2. Generate Core Artifacts: `paper.tex` and `audit_claims.py` The Reproducibility dimension (score: 20) failed because foundational artifacts were missing. The paper draft serves as the canonical summary of claims, methods, and results; without it, there is nothing to audit. The `audit_claims.py` script ensures that every quantitative claim in the paper can be traced to source data. The next run should treat `paper.tex` generation as a mid-run checkpoint, not a final deliverable. Draft early, audit continuously.

3. Execute Domain-Specific Disclosures Explicitly The Domain Depth dimension (score: 30) failed across all three probes: no explicit active-space statement, no fermion-to-qubit mapping reference, and no simulator precision floor disclosure. These are not stylistic preferences—they are minimum credibility requirements for quantum chemistry claims. The next run should include a `domain_disclosures.md` template that forces explicit answers to these questions before any simulation executes. This prevents downstream reviewers from dismissing results due to unstated assumptions.

Bottom line: This run produced no defensible scientific output. The path forward requires treating artifact generation as mandatory infrastructure, not optional polish.