QuantumNovelty Review Showcase

Paper audit Quantum Convolutional Neural Networks are Effectively Classically Simulable (arXiv:2408.12739)

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
`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

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)
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**
Grand total		**3,008**	**9,896**	**$1.3591**	**233.4**

3 Verdict summary

Paper	Panel score	EIC verdict	CQE composite
`qcnn`	6.0/10	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 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
PRX Quantum 7, 020304 (2026)

4.1 qcnn — Deep-research review

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

4.1.1 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.

4.1.2 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 (_abs, _rel); 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 `audit_claims.py` 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.

4.1.3 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

4.1.4 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.

4.2 qcnn — Reviewer panel 5 voices

Model: `claude-opus-4-5-20251101` · 4,932 output tokens · 121.7 s

4.2.1 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 ~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.

4.2.2 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.

4.2.3 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.

4.2.4 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.

4.2.5 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.

4.2.6 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

4.3 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 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 ś 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 ś standard deviation across all 5 runs for each dataset and encoding combination, not just the best single run.

4.4 qcnn — Stage-6 CQE narrative

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

Process Summary: QuantumNovelty Run Evaluation

4.4.1 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.

4.4.2 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.

4.4.3 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.

4.4.4 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.mo` `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.