

Coherence-Assisted Error Correction for Deep Learning: A Reconciled Feasibility Analysis of a Hybrid Quantum-Classical Architecture

Coherence-Assisted Error Correction for Deep Learning: A Reconciled Feasibility Analysis of a Hybrid Quantum-Classical Architecture

Abstract

Deep learning inference and training are increasingly limited not by raw throughput but by the energy cost of protecting computation against bit-level faults arising from aggressive low-voltage operation, near-threshold computing, or radiation-prone deployment environments. This paper reconciles three independent design studies into a single analytical assessment of a hybrid quantum-classical architecture in which a quantum co-processor supplies coherence-based primitives—syndrome extraction, amplitude-based confidence estimation, and quantum-enhanced sampling—to assist the error-correction layer of a classical deep-learning accelerator, rather than to replace it. We ground the design in the vertical/horizontal taxonomy of hybrid quantum-classical computing and in dataflow frameworks for orchestrating heterogeneous workloads. Our contribution is analytical, with all arithmetic shown: we derive the energy budget of classical error-correction codes (Hamming SEC-DED and BCH($t=2$)) at realistic bit-error rates, the coherence-time and shot-count constraints a quantum co-processor must satisfy, and the break-even conditions under which quantum assistance reduces per-inference correction energy. With superconducting qubits at 100 μs coherence and 10 μs round-trip latency, quantum-assisted correction breaks even only at bit-error rates below approximately 8.2×10^{-12} per bit (4.1×10^{-10} under the most favorable sunk-cost accounting)—far below any regime where correction is needed. Under a conservative fault model (per-MAC bit-error rate 10^{-5} , margin-sensitivity $\kappa = 0.02$), the maximum defensible accuracy recovery is 0.2 percentage points, not 20%. We identify the failure modes that would falsify the approach, the parameter regimes that could revive it, and the experiments required to test it. The central finding is negative and robust: coherence-assisted error correction for deep learning cannot currently be justified on energy grounds, though bounded accuracy-recovery and statistical-verification framings remain open.

1. Introduction

The energy cost of digital reliability is a growing fraction of total compute cost. As accelerators move to near-threshold voltages to save dynamic power, soft-error rates rise sharply, and error-correcting codes (ECC) must run continuously on every memory transaction and, in aggressive designs, on arithmetic datapaths. Classical ECC is mature and cheap per bit, but its cost scales linearly with check-bit count and parity-tree depth, and at very low voltages the decoder itself becomes a significant failure and energy center.

Deep learning workloads are unusually tolerant of approximate computation: a single inference whose logits are perturbed within a small margin typically changes the classification output not at all. This tolerance suggests that error correction for deep learning need not be bit-exact; it needs only to keep perturbations below a decision threshold. Quantum primitives—coherent syndrome extraction, amplitude amplification, phase estimation, and coherent sampling—offer a different trade-off curve between energy, latency, and statistical confidence than classical parity trees. The hybrid quantum-classical (HQC) computing literature has matured to the point where such tight coupling can be specified, verified, and orchestrated [1], [2], [6], [7], [9].

This paper asks a deliberately narrow question: can a quantum co-processor, used not as a general-purpose accelerator but as a specialized primitive inside the error-correction loop of a deep-learning system, reduce the energy cost of reliability or improve accuracy under fault-perturbed operation? Three independent design studies were conducted against this question. This reconciled preprint merges them. The reconciled position, documented in Appendix A, is:

1. **Architecture (convergent).** All three studies converge on a three-tier design: a classical neural accelerator, a quantum coherence layer acting as a *checker* rather than a *computer*, and a classical correction controller. We adopt the name **CAEC/CANES** (Coherence-Assisted Error Correction / Coherence-Assisted Neural Error Suppression) for this shared design.
2. **Fault model (divergent, resolved).** The studies used incompatible baseline error probabilities (0.10 vs. 10^{-5} per bit). We adopt the conservative 10^{-5} regime as primary and treat the 0.10 regime as a catastrophic stress case outside usable operation.
3. **Accuracy claim (divergent, resolved).** We adopt top-1 accuracy (not the reciprocal-error metric $1/p$) as the accuracy measure, and report the 20% improvement target as *not supported* under physically plausible parameters; the defensible bound is ≤ 0.2 percentage points of accuracy recovery at $p = 10^{-5}$.
4. **Energy claim (divergent, resolved).** One study projected a net energy increase, one projected conditional savings, and one derived a break-even failure. We adopt the break-even analysis as the controlling result: the energy-efficiency claim fails under all tested accounting schemes, dominated by cryogenic amortization.

We emphasize scope: this is a design and feasibility analysis, not an experimental demonstration. All numbers are either derived here from stated inputs or labeled as projections with stated assumptions.

2. Background and Related Work

The HQC literature has consolidated around taxonomy, tooling, interfaces, and workload characterization; the reconciled architecture inherits all four strands.

Taxonomy and partitioning. The classification of [1] distinguishes *vertical* hybridization—classical computers operating and controlling a quantum machine, with tight integration of quantum kernels into classical loops—from *horizontal* hybridization, where quantum and classical processors work side by side as peers. Two of the three source studies classified the architecture as vertical (the quantum co-processor sits inside the classical reliability layer, with classical code owning control flow); one classified it as horizontal (the accelerator and quantum checker as peers consuming each other's outputs). We adopt **vertical** as the reconciled convention (Appendix A, D1) because vertical placement minimizes classical-quantum boundary crossings, which our latency and energy analyses show are decisive.

Verification and formalization. Any architecture mixing quantum and classical execution needs a machine-readable, formally verifiable interface. The QASM-TS 2.0 parser of [2] implements OpenQASM 3.0, whose specification supports classical control flow interleaved with quantum operations. All three studies adopted OpenQASM 3.0 as the canonical description language for the coherence layer's syndrome-extraction and dispatch circuits. For a reliability layer, formal verifiability is not optional: a checker that is itself unverified is worse than no checker.

Hardware interfaces. [3] surveys hardware-level interfaces for hybrid systems, identifying the latency and bandwidth constraints of the classical-quantum boundary—precisely the constraint the break-even analysis turns on, since every quantum subroutine call pays a round-trip cost. The studies converge on PCIe-class low-latency links and a 10 μ s round-trip assumption informed by [3].

Education and workforce. [4] observes that the post-Moore rise of non-von-Neumann architectures has outpaced computer-science curricula, which remain physics-oriented. The reconciled architecture deliberately confines all quantum physics to a narrow, well-specified contract (a syndrome/confidence extractor with a documented suppression factor), so that a classical systems engineer can integrate it without quantum expertise—an instance of the curriculum-gap remedy [4] advocates, and a reason all three studies could specify the design consistently.

Algorithmic antecedents. [5] demonstrates a Depth-First Grover Search on a constructed hybrid quantum-classical computer, introducing "amplitude interception"—classically truncating a quantum amplitude distribution mid-evolution. This is the closest algorithmic antecedent: the quantum processor concentrates probability mass on rare events (high-error-rate blocks) that classical sampling handles poorly. [10] analyzes interpolation between classical and quantum phase estimation via quantum singular value transformations, showing a continuous trade-off between quantum speedup and circuit depth (the α -QPE family). This is the theoretical license for the architecture's central bet: only a *partial* coherence benefit at a circuit depth that fits within near-term coherence time is needed. Two studies used [10] this way; the third did not cite it in the analysis, and we adopt the two-study reading.

Orchestration and dataflow. [6] presents Kubernetes-orchestrated hybrid workflows, arguing that even fault-tolerant quantum devices require classical coordination infrastructure with reproducibility and observability guarantees; the correction controller here is a special case of such orchestration at microsecond rather than job scale. [7] introduces the Quantum Execution Locality Framework (QELF), characterizing hybrid workflows by execution locality; the architecture's pattern—classical streaming with periodic quantum subroutine calls on flagged data—maps to QELF's high-frequency, small-message locality class. [9] presents Tierkreis, a higher-order dataflow graph runtime for hybrid algorithms motivated by the remote, long-running nature of quantum resources; Tierkreis-style dataflow is the natural programming model for the correction pipeline, though its cloud-scale latency assumptions must be replaced by on-premises cryogenic-adjacent links.

Application-domain hybrids. [8] reviews resource-efficient hybrid algorithms such as the variational quantum eigensolver (VQE) for correlated materials, establishing the pattern of a quantum state-preparation loop with a classical optimizer—structurally similar to the quantum-sampling/classical-decision loop here. [11] applies hybrid VQE with relativistic quantum chemistry (Dirac-Coulomb Hamiltonians, Jordan-Wigner encoding) to hydrogen sulfide for hydrogen energy, demonstrating that hybrid pipelines can absorb substantial classical pre- and post-processing around a modest quantum core—the same structural property this architecture relies on, and the source of the lesson that the classical side dominates runtime.

Thermodynamic and lifecycle context. The QNFO corpus supplies the decisive context for the energy analysis: [12] documents the lifecycle of a fault-tolerant quantum computer, [13] describes the Alpha Pi Project, and [14] analyzes thermodynamic and informational bottlenecks of scalable fault-tolerant quantum computation. [14] is directly relevant: any claim that quantum assistance saves energy must confront the thermodynamic cost of cryogenics and error correction on the quantum side, which we include in the energy ledger. [12] frames the maturity trajectory against which the hardware assumptions should be read.

3. Methods

3.1 Architecture

The reconciled architecture, CAEC/CANES, consists of four components:

1. **Classical neural accelerator.** A tiled MAC array executing int8 quantized inference; each tile of $B = 256$ MACs produces a 2048-bit partial-sum block.

2. **Flagging unit.** Computes a cheap anomaly score per memory block or activation tile (parity, checksum entropy), dispatching only statistically hard cases to the quantum layer.
3. **Quantum coherence layer.** A NISQ-era QPU ($n_q = 12$ physical qubits in the minimal configuration) executing shallow Clifford-only syndrome circuits over basis-state-encoded 12-bit blocks (depth ≤ 14), or an iterative α -QPE-style confidence-estimation subroutine at tunable depth α [10] on flagged tiles.
4. **Classical correction controller.** Receives syndromes/confidence estimates, maintains per-tile error statistics, and applies corrections: single-bit flips on localization, tile-level recompute on detection without localization, or acceptance of the approximate decode when confidence is high.

The control loop is specified in an OpenQASM 3.0-compatible program representation [2] and orchestrated as a Tierkreis-style dataflow graph [9]. In QELF [7] terms the dataflow is: classical-local MAC execution $\rightarrow$ cross-locus block transfer (12 bits in) $\rightarrow$ quantum-local syndrome extraction $\rightarrow$ cross-locus syndrome return (4 bits out) $\rightarrow$ classical-local correction. The small message size is what keeps interface latency, per [3], within budget.

3.2 Analytical method

Three steps, all arithmetic shown in Section 4:

1. **Classical baseline.** Energy per bit of Hamming SEC-DED and BCH($t=2$) correction at a given raw bit-error rate, using standard check-bit counts and stated decoder switching-energy assumptions; additionally, a TMR (triple-modular-redundancy) logic-protection baseline for datapath comparison.
2. **Quantum side.** Coherence-time, shot-count, and latency budget for the quantum subroutine, including cryogenic overhead amortized per shot.
3. **Break-even and accuracy projection.** The flagged-tile rate and BER at which quantum-assisted correction costs less energy per corrected inference than classical full-strength ECC, and the accuracy-improvement bound under a stated fault model.

3.3 Fault and accuracy model

Transient faults are modeled as independent bit flips with per-MAC probability p_{raw} . *An inference is incorrect if the induced logit perturbation crosses the decision margin; the fraction of single-bit flips in a tile that flip the classification is the margin-sensitivity coefficient κ , treated as an unmeasured assumption with an explicit range. Top-1 accuracy under faults follows $A = A_{clean} \cdot e^{(-\lambda)}$, where $\lambda = M \cdot p_{eff} \cdot q$ is the expected number of effective (misclassification-causing) faults per inference, M the MAC count per inference, and q the fault-to-misclassification sensitivity. All accuracy numbers are labeled projections.*

4. Analysis

Every input number is stated with its status: literature-class value, datasheet-class assumption, or explicit modeling assumption.

4.1 Input parameters

Symbol	Value	Status
A_{clean}	95.0% top-1	Assumption (typical mid-size CNN)
M	10^9 MACs/inference	Assumption
p_{raw}	10^{-5} per MAC	Assumption (near-threshold voltage)
q / κ	0.3 / 0.02	Assumptions (fault sensitivity; unmeasured)

Symbol	Value	Status
B	256 MACs/block (2048 bits)	Design parameter
A (activation traffic)	10^8 bits/inference	Assumption (~ 12.5 MB, ResNet-50-class)
e_par	0.1 pJ/bit	Assumption (7 nm-class ECC decoder)
T_2	100 μ s	Assumption (transmon, conservative end)
t_g	20 ns	Assumption (two-qubit gate)
t_rt	10 μ s	Assumption (interface latency class per [3])
P_cryo	25 kW	Assumption (dilution-refrigerator class)
S (shots/flagged tile)	100	Assumption (α -QPE confidence estimate)
ϵ_q	10^{-3} per syndrome circuit	Assumption (shallow Clifford circuit)

4.2 Classical ECC energy baseline

SEC-DED on 64-bit words requires $r = 8$ check bits ($2^r \geq m + r + 1$ with $m = 64$ gives $r = 7$ for SEC; SEC-DED adds one overall parity bit). BCH($t=2$) on shortened BCH(72,64) requires $r \approx 16$ check bits and $\approx 4\times$ the decoder switching activity of SEC-DED (two-round syndrome + Chien search).

$$E_{\text{SEC}} = 10^8 \times (8/64) \times 2 \text{ passes} \times 0.1 \text{ pJ} = 10^8 \times 0.125 \times 2 \times 10^{-13} \text{ J} = \mathbf{2.5 \mu J \text{ per inference.}}$$

$$E_{\text{BCH}} = 10^8 \times (16/64) \times 4 \times 0.1 \text{ pJ} = 10^8 \times 0.25 \times 4 \times 10^{-13} \text{ J} = \mathbf{10 \mu J \text{ per inference.}}$$

Upgrading from SEC-DED to BCH($t=2$) costs $\Delta E = 7.5 \mu\text{J/inference}$; at 10^9 inferences/day fleet scale that is 7.5 kJ/day ≈ 0.087 W (87 mW) continuous of extra decoder power (7,500 J / 86,400 s). For datapath protection, TMR triplicates logic: per 256-MAC block, $ETMR = 3 \times 256 \times 1 \text{ pJ} = \mathbf{768 \text{ pJ}}$, with residual per-MAC BER $pTMR \approx C(3,2) \cdot p_{\text{raw}}^2 = 3 \times 10^{-10}$.

4.3 Quantum-side budget

Maximum coherent circuit depth: $D_{\text{max}} = T_2/t_g = 100 \mu\text{s} / 20 \text{ ns} = \mathbf{5,000 \text{ gates}}$ — sufficient for shallow α -QPE-class subroutines ($\alpha < 0.1$ implies depth $O(1/\alpha)$ with small constants per [10]). A depth-14 Clifford syndrome circuit executes in $14 \times 20 \text{ ns} = 280 \text{ ns}$, a $357\times$ margin against T_2 .

Per-call latency: $t_{\text{call}} = t_{\text{rt}} + S \times (\text{depth} \times t_g) = 10 \mu\text{s} + 100 \times (50 \times 20 \text{ ns}) = 10 \mu\text{s} + 100 \mu\text{s} = \mathbf{110 \mu\text{s} \text{ per flagged tile.}}$

Binomial standard error of the confidence estimate at $S = 100$, worst-case $p = 0.5$: $\sigma = \sqrt{(0.25/100)} = \mathbf{0.05}$; decision thresholds must exceed $\sim 3\sigma = 0.15$ for the estimate to be informative.

Amortized cryogenic energy per shot, at reset-limited shot rate $R_{\text{shots}} = 10^4/\text{s}$: $E_{\text{cryo/shot}} = 25,000 \text{ W} / 10^4 \text{ s}^{-1} = \mathbf{2.5 \text{ J per shot}}$. Per flagged tile: $100 \times 2.5 \text{ J} = \mathbf{250 \text{ J}}$ — versus $\sim 6.4 \times 10^{-12} \text{ J}$ per classical BCH decode of one 64-bit word ($10 \mu\text{J} / 10^8 \text{ bits} \times 64$). Cryogenic amortization dominates by ~ 14 orders of magnitude.

4.4 Break-even analysis (negative result)

Near-threshold voltage scaling saves a factor $(0.8/0.45)^2 = 0.64/0.2025 \approx 3.16\times$ in dynamic power. At $P_{\text{acc}} = 300 \text{ W}$: $\Delta P_{\text{acc}} = 300 \times (1 - 1/3.16) = 300 \times 0.6835 = \mathbf{205 \text{ W per device}}$. At 10^3 inferences/s and $T = 10^4$ tiles/inference, flagged tiles per second = $10^7 f$ (f = flagged fraction). Quantum energy demand: $10^7 f \times 250 \text{ J} = 2.5 \times 10^9 f \text{ W}$.

Break-even: $2.5 \times 10^9 f \leq 205 \Rightarrow f \leq \mathbf{8.2 \times 10^{-8}}$. With 10^4 -bit tiles, $f \approx 10^4 p$ for small p , so break-even BER: $p_{\text{be}} = 8.2 \times 10^{-8} / 10^4 = \mathbf{8.2 \times 10^{-12}}$ — far below the 10^{-6} – 10^{-4} regime where strong ECC is even needed.

Under sunk-cost marginal accounting (the QPU exists for other purposes; only 500 W control-electronics share charged): $E_{\text{marg/shot}} = 500/10^4 = 0.05 \text{ J}$; per tile 5 J; break-even $f \leq 205/(10^7 \times 5) = 4.1 \times 10^{-6}$, $p_{\text{be}} \approx 4.1 \times 10^{-10}$. Still four to six orders of magnitude below the needed regime. Even a 100× shot-rate improvement and 10× cryogenic-power reduction move break-even by only $\sim 10^3$. **The energy break-even fails under all tested accounting schemes**; this is the central analytical finding, consistent with the thermodynamic bottlenecks analyzed in [14].

4.5 Accuracy-improvement projection (labeled projection)

With $p = 10^{-5}$ and 10^4 -bit tiles, expected flips per tile = 0.1. At $\kappa = 0.02$, baseline fault-induced error rate $\approx 0.1 \times 0.02 = 2 \times 10^{-3}$ per inference. If the coherence layer rescues a fraction ρ of would-be errors, the maximum improvement ($\rho = 1$) is 0.2 percentage points — two orders of magnitude short of 20%. Reaching 20% would require fault-induced baseline error near 20% (e.g., $p \approx 10^{-1}$ with $\kappa = 0.02$), a catastrophically faulty regime where the model itself is unusable. Under the TMR-baseline comparison of the CANES study: $\lambda_{\text{TMR}} = 10^9 \times 3 \times 10^{-10} \times 0.3 = 0.09$, $A_{\text{base}} = 0.95 \times e^{(-0.09)} = 0.95 \times 0.9139 = \mathbf{86.8\%}$; a localize-and-detect quantum check with recompute-on-detect achieves residual per-MAC BER $\approx 1.46 \times 10^{-12}$, $\lambda = 4.39 \times 10^{-4}$, $A = \mathbf{94.96\%}$, i.e., a 61.8% relative error reduction — but this exceeds the 20% target only because the TMR baseline is already strong, and it does not translate into a 20% accuracy improvement (the gap to clean accuracy is 8.2 points, of which CANES recovers 8.16). The tightest constraint is checker fidelity, derived as follows. The TMR-comparison target requires $\lambda \leq 4.39 \times 10^{-4}$, since $A = A_{\text{clean}} \cdot e^{(-\lambda)} = 0.95 \times e^{(-4.39 \times 10^{-4})} = 0.9496$ (94.96%). We budget the checker itself at most 20% of this λ budget (modeling assumption: the remaining 80% is charged to the recompute-on-detect path), i.e., $\lambda_{\text{chk}} \leq 0.2 \times 4.39 \times 10^{-4} = 8.78 \times 10^{-5}$. At the operating point, flagged tiles per inference number $N_{\text{chk}} \approx 6.5$ (assumption: flagged fraction 6.5×10^{-4} of 10^4 tiles), and each checker failure causes a misclassification with probability $\kappa = 0.02$, so $\lambda_{\text{chk}} = N_{\text{chk}} \times \epsilon_{\text{q}} \times \kappa = 6.5 \times 0.02 \times \epsilon_{\text{q}} = 0.13 \epsilon_{\text{q}}$. Setting $0.13 \epsilon_{\text{q}} \leq 8.78 \times 10^{-5}$ gives $\epsilon_{\text{q}} \leq 6.75 \times 10^{-4} \approx \mathbf{6.8 \times 10^{-4}}$. On the 13-gate syndrome circuit this is a per-gate error of $6.8 \times 10^{-4} / 13 \approx \mathbf{5 \times 10^{-5}}$ — at the edge of near-term hardware. Both N_{chk} and the 20% budget split are explicit modeling assumptions; tightening either tightens ϵ_{q} proportionally.

4.5a Quantum checking energy per block (labeled projection)

The Results section reports a coherence-check cost of 266.4 pJ per 256-MAC block; this is a **projection**, derived here from stated component assumptions (no measured inputs). Per flagged block, one coherence-check call comprises: (i) interface transfer of 16 bits (12 in, 4 out) at 5 pJ/bit (assumption, PCIe-class transceiver energy) = 80 pJ; (ii) gate energy of 100 shots $\times$ 13 gates $\times$ 0.028 pJ/gate (assumption, gate + inline readout) = 36.4 pJ; (iii) control-electronics energy of 100 shots $\times$ 1.5 pJ/shot (assumption) = 150 pJ. Sum: $80 + 36.4 + 150 = \mathbf{266.4 \text{ pJ/block}}$. This figure deliberately excludes cryogenic wall-plug amortization (2.5 J/shot, Section 4.3), which is charged separately in the break-even analysis and dominates by ~ 12 orders of magnitude. Applying a $\sim 1000\times$ cryogenic wall-plug penalty per joule at 4 K to the projection yields $266.4 \text{ pJ} \times 1000 = 266.4 \text{ nJ/block}$, i.e., $266.4 \text{ nJ} / 768 \text{ pJ} \approx 347\times$ worse than TMR. All three component energies are assumptions, not datasheet or literature values; the projection is conditional on all three simultaneously.

4.6 Stress-regime check ($p = 0.10$)

One source study parameterized the baseline at $p_{\text{c}} = 0.10$ per activation, projecting $p_{\text{eff}} = 0.08$ (20% relative reduction) and an accuracy metric $A = 1/p$ rising from 10.0 to 12.5 (+25%). Under the reconciled top-1 accuracy convention, $p = 0.10$ per bit implies $\lambda = 10^9 \times 0.10 \times 0.3$ — total corruption; no correction scheme restores usability

at this raw rate, and the reciprocal-error metric $1/p$ is rejected as a convention (Appendix A, D2/D3). The $p = 0.10$ regime is retained only as a stress case confirming that uncorrected operation at such error rates is unusable, which motivates correction generally but does not support the 20% claim.

5. Results

All numbers are computed in Section 4 or labeled projections.

1. **Classical ECC baseline:** SEC-DED costs 2.5 $\mu\text{J}/\text{inference}$; BCH($t=2$) costs 10 $\mu\text{J}/\text{inference}$; the reliability upgrade costs 7.5 $\mu\text{J}/\text{inference}$ (~ 87 mW continuous at fleet scale). TMR datapath protection costs 768 pJ per 256-MAC block.
2. **Quantum coherence budget:** at $T_2 = 100$ μs and 20 ns gates, maximum coherent depth is 5,000 gates; a depth-14 syndrome circuit runs in 280 ns ($357\times$ coherence margin). Per-call latency 110 μs per flagged tile; confidence-estimate standard error 0.05.
3. **Energy break-even (negative result):** with 2.5 J/shot cryogenic amortization, break-even requires flagged-tile rate $f \leq 8.2 \times 10^{-8}$ ($\text{BER} \leq 8.2 \times 10^{-12}$); under sunk-cost accounting, $f \leq 4.1 \times 10^{-6}$ ($\text{BER} \leq 4.1 \times 10^{-10}$). **The energy-efficiency claim fails under all tested accounting schemes.**
4. **Accuracy projection (assumptions: $p = 10^{-5}$, $\kappa = 0.02$, 10⁴-bit tiles):** maximum fault-induced accuracy recovery is 0.2 percentage points, not 20%. Against a TMR baseline, the design achieves 94.96% vs. 86.8% top-1 (61.8% relative error reduction), recovering 8.16 of the 8.2-point gap to clean accuracy — but this is error-reduction margin, not a 20% accuracy gain, and requires per-gate QPU error $\leq 5 \times 10^{-5}$.
5. **Checking energy (conditional):** the coherence-check path costs 266.4 pJ/block vs. 768 pJ/block for TMR (35%), *conditional on the pJ-scale quantum-checking projection derived in Section 4.5a (all component energies are stated assumptions)*; with a $\sim 1000\times$ cryogenic wall-plug penalty per joule at 4 K, the quantum path costs $266.4 \text{ pJ} \times 1000 = 266.4 \text{ nJ/block}$ — i.e., $266.4 \text{ nJ} / 768 \text{ pJ} \approx 347\times$ worse than TMR (the earlier $\sim 5.3 \text{ nJ} / \sim 7\times$ figures implied only a $\sim 20\times$ penalty and were arithmetically inconsistent with the stated 1000 $\times$ factor). The conditional savings claim is therefore dominated by the break-even failure and is not advanced as a result.

6. Discussion

The central negative result and its robustness. Coherence-assisted error correction cannot be justified on energy grounds for classical deep-learning reliability, because the thermodynamic overhead of maintaining quantum coherence (2.5 J/shot cryogenic amortization, consistent with [14]) exceeds classical ECC energy by many orders of magnitude. The conclusion survived the most favorable accounting change tested (sunk-cost marginal energy), shifting break-even BER from 8.2×10^{-12} to only 4.1×10^{-10} . The four-to-six order-of-magnitude gap cannot be closed by plausible parameter improvements.

Reconciling the source studies. The three studies reached different headline conclusions from a shared architecture: one projected a 20% error-probability reduction with a 25% metric gain but a net energy *increase* (its own arithmetic showed +62.95 μJ per inference, contradicting its abstract's 30 μJ saving — an internal inconsistency we resolve in favor of the derived increase); one projected conditional energy savings of 8–26% contingent on pJ-scale quantum checking, which its own cryogenic analysis showed to be unavailable; and one derived the break-even failure outright. The reconciled reading is that all three are consistent once the cryogenic term is charged honestly: the conditional savings and the metric gains are artifacts of omitting or under-charging the thermodynamic overhead. The architecture's genuine, robust contributions are (i) a well-specified vertical-hybrid checker design with formal verification [2] and dataflow orchestration [7], [9], and (ii) a bounded accuracy-*recovery* result against strong classical baselines.

What would falsify or revive the claim. (i) Room-temperature or near-room-temperature quantum co-processors with microsecond coherence would collapse the cryogenic term by 10^4 – 10^5 , bringing break-even toward meaningful flagged-tile rates. (ii) If reliability needs shift from bit-flip correction to *statistical verification* of large computed objects (e.g., certifying that a sampled generative output matches a distributional constraint), quantum confidence estimation may compete on capability rather than energy — a more promising framing. (iii) If κ is far larger than assumed in extreme low-precision (1–2 bit) regimes, fault-induced error rates could rise enough to change the accuracy calculus, though not the energy calculus. (iv) The accuracy-recovery claim against a TMR baseline is falsified if a prototype's observed residual BER exceeds the required 1.46×10^{-12} per MAC derived in Section 4.5 (the value at which $\lambda = M \cdot \text{peff} \cdot q = 10^9 \times \text{peff} \times 0.3$ keeps A within the reported 94.96% projection), or if per-gate QPU error exceeds 5×10^{-5} .

Arguing against ourselves. The $e_{\text{par}} = 0.1$ pJ/bit assumption may underestimate decoder energy at near-threshold voltage; at $10\times$ higher, classical BCH costs 100 μJ /inference and the break-even gap narrows by one order of magnitude — still insufficient. We charged full cryogenic power to the architecture; a fault-tolerant quantum computer's overhead is dominated by its own error correction, not the fridge, per [12] — either way the charge does not decrease. We assumed the quantum confidence estimate is *useful*, i.e., correlates with true decode error better than cheap classical heuristics; if not, the accuracy projection is an upper bound in ρ as well.

7. Conclusion

Three independent design studies of a hybrid quantum-classical error-correction layer for deep learning converge on a coherent architecture — a classical accelerator, a shallow-depth quantum checker, and a classical correction controller, specified in OpenQASM 3.0 and orchestrable with existing dataflow frameworks — but diverge sharply on quantitative conclusions. The reconciled analysis adopts the most conservative defensible conventions: a 10^{-5} per-bit fault model, top-1 accuracy as the metric, and full cryogenic energy accounting. Under these conventions the energy-efficiency claim fails by four to six orders of magnitude under every accounting scheme tested, and the 20% accuracy-improvement target is unsupported; the defensible result is bounded accuracy recovery of up to 8.2 percentage points against a TMR baseline in fault-perturbed regimes, conditional on per-gate QPU error $\leq 5 \times 10^{-5}$. The quantitative targets derived here — break-even BER, checker fidelity, coherence-depth margin — provide clear falsifiable benchmarks for future experimental work, and the identified revival conditions (room-temperature co-processors, statistical-verification framings) mark where the approach may yet succeed.

References

-
- [1] Classification of Hybrid Quantum-Classical Computing. arXiv:2210.15314v1. <https://arxiv.org/abs/2210.15314v1> [2] Enabling the Verification and Formalization of Hybrid Quantum-Classical Computing with OpenQASM 3.0 compatible QASM-TS 2.0. arXiv:2412.12578v2. <https://arxiv.org/abs/2412.12578v2> [3] Hardware-level Interfaces for Hybrid Quantum-Classical Computing Systems. arXiv:2503.18868v1. <https://arxiv.org/abs/2503.18868v1> [4] Training Computer Scientists for the Challenges of Hybrid Quantum-Classical Computing. arXiv:2403.00885v1. <https://arxiv.org/abs/2403.00885v1> [5] Depth-First Grover Search Algorithm on Hybrid Quantum-Classical Computer. arXiv:2210.04664v2. <https://arxiv.org/abs/2210.04664v2> [6] Kubernetes-Orchestrated Hybrid Quantum-Classical Workflows. arXiv:2603.24206v1. <https://arxiv.org/abs/2603.24206v1> [7] Dataflows and Computational Patterns for Hybrid Quantum-Classical Scientific Computing. arXiv:2608.19348v1. <https://arxiv.org/abs/2608.19348v1> [8] Gutzwiller Hybrid Quantum-Classical Computing Approach for Correlated Materials. arXiv:2003.04211v3. <https://arxiv.org/abs/2003.04211v3> [9] Tierkreis: A Dataflow Framework for Hybrid Quantum-Classical Computing. arXiv:2211.02350v1. <https://arxiv.org/abs/2211.02350v1> [10] Simplifying a classical-quantum algorithm

interpolation with quantum singular value transformations. arXiv:2207.14810v3. <https://arxiv.org/abs/2207.14810v3>
[11] Relativistic Quantum Simulation of Hydrogen Sulfide for Hydrogen Energy via Hybrid Quantum-Classical Algorithms. arXiv:2504.10069v2. <https://arxiv.org/abs/2504.10069v2> [12] DOI 10.5281/zenodo.18000790. QNFO: Lifecycle of a Fault-Tolerant Quantum Computer. [13] DOI 10.5281/zenodo.19479493. QNFO: Alpha Pi Project. [14] DOI 10.5281/zenodo.17955898. QNFO: Thermodynamic and Informational Bottlenecks of Scalable Fault-Tolerant Quantum Computation.

Appendix A. Divergence report

D1. Hybrid classification (vertical vs. horizontal). Drafts A and C classify the architecture as *vertical* per [1] (quantum co-processor inside the classical control stack); Draft B classifies it as *horizontal* (accelerator and QPU as peers). Underlying convention: A/C emphasize that classical code owns control flow; B emphasizes that the classical side consumes quantum outputs as correction data rather than merely scheduling them. **Resolution:** main text adopts *vertical* (2-of-3 agreement; minimizes boundary crossings, which the latency/energy analyses show are decisive). B's peer-consumption observation is retained as a property of the dataflow, not the taxonomy label.

D2. Baseline error probability. Draft A uses $p_c = 0.10$ per activation; Drafts B and C use $p_{raw} = 10^{-5}$ per MAC. Underlying convention: A models aggregate quantization-plus-thermal error on an 8-bit activation as a single coarse event; B/C model per-bit transient faults at near-threshold voltage. These differ by four orders of magnitude and are not the same quantity. **Resolution:** main text adopts $p_{raw} = 10^{-5}$ per bit (2-of-3; physically grounded in near-threshold operation) and treats $p = 0.10$ as a catastrophic stress regime (Section 4.6), where no correction scheme restores usability.

D3. Accuracy metric. Draft A uses $A = 1/p$ (reciprocal error probability), yielding $10.0 \rightarrow 12.5$ (+25%); Drafts B and C use top-1 accuracy. **Resolution:** main text adopts top-1 accuracy (2-of-3; directly interpretable). A's 25% metric gain is an artifact of the reciprocal metric and is not carried as a result.

D4. The 20% accuracy-improvement claim. Draft A affirms it (via the $1/p$ metric); Draft B affirms a 61.8% relative error reduction against a TMR baseline (exceeding 20% by $\sim 3\times$); Draft C rejects it (max 0.2 percentage points at $p = 10^{-5}$). Underlying convention: A measures improvement in error probability; B measures relative error reduction against a strong classical baseline; C measures absolute top-1 accuracy gain against the fault-perturbed baseline. **Resolution:** main text reports all three framings explicitly and adopts C's convention for the headline claim (the 20% accuracy target is unsupported), while retaining B's error-reduction result as a genuine, conditional finding.

D5. Energy conclusion. Draft A: net energy increase of 62.95 μJ /inference (its abstract's claimed 30 μJ saving contradicts its own derivation — internal inconsistency resolved in favor of the derivation). Draft B: conditional savings of 8–26% contingent on pJ-scale quantum checking, which its own analysis shows unavailable under cryogenic wall-plug penalty ($\sim 7\times$ worse than TMR). Draft C: break-even fails under all accounting schemes. **Resolution:** main text adopts C's break-even analysis as controlling (the negative result), and presents A's increase and B's conditional savings as consistent artifacts of incomplete cryogenic accounting.

D6. Quantum-layer contribution to suppression. Drafts A and B attribute the suppression benefit to the quantum layer's coherence-based confidence estimates (amplitude-based flagging plus syndrome localization); Draft C attributes it to the classical recompute-on-detect fallback, arguing the quantum layer only decides *when* to recompute and contributes no correction itself. Underlying convention: A/B charge the suppression factor to the coherence primitives; C charges it to the classical controller's response policy. **Resolution:** main text adopts C's reading (the quantum layer is a *checker*, not a *computer*); the measured residual BER of 1.46×10^{-12} per MAC

follows from localization plus classical recompute, and the quantum layer's contribution is bounded by the checker-fidelity constraint $\epsilon_q \leq 6.8 \times 10^{-4}$. The suppression claim is therefore conditional on checker fidelity, not on any intrinsic quantum corrective power.