AIRR:ARR-2026-79N8QC5BN98F283V v1 [Quantum physics] 27 Nov 2026 · AIRR.SCIENCE

Exact Bloch-Sphere Surrogates, Heisenberg R_{zz} Entangler Leakage, and Token-Mixing Barriers for Multi-Head Quantum Neural Networks inside Large Language Models

Lluis Eriksson*

November 27, 2026

Abstract

Amoussou, Aizpurua, Singh, Orús, and Jacobsen [1] recently demonstrated that a single multi-head quantum neural network (QNN) with 145,281 trainable parameters ($D = 2048$, $k = 32$ heads, $n_q = 8$ qubits, $L = 4$ layers) can replace up to eight contiguous transformer blocks inside Qwen3-30B-A3B and execute on a 156-qubit IBM Heron processor. Their ablation study uncovered three striking empirical anomalies left open theoretically: (i) deleting the quantum heads leaves a mixer-only control $M_2 \tanh(M_1 x)$ whose residual gate collapses ($\alpha = 0.0042$–0.0056 on two of three seeds) and loses 3.5 perplexity points on sensitive Layer 1; (ii) switching all R_{zz} entanglers off or freezing circuit angles consistently *beats* the trained entangled circuit on held-out perplexity; and (iii) replacing scattered sensitive blocks fails to compose and degrades long-range context tasks (LAMBADA). We resolve all three anomalies algebraically from first principles, backed by 20 constructive Lean 4.36.0 theorems and 16 zero-float $\mathbb{Q}[i]$ certificates. First, when entanglers vanish ($\gamma = 0$), every n_q-qubit, L-layer dense-angle-encoded head collapses identically to n_q decoupled SO(3) Bloch evolutions, yielding an exact $O(n_q L)$ classical surrogate reproducing all 1- and 2-qubit Pauli expectations with $2 \cdot 2^{n_q}/(3n_q) = 64/3\times$ state compression at $n_q = 8$ and zero 2^{n_q} statevector cost. Second, we prove the exact *odd-parity collapse* of the mixer-only control ($g_{\mathrm{mix}}(-x) = -g_{\mathrm{mix}}(x) \implies \mathrm{MSE}(f, g_{\mathrm{mix}}) \geq \|f_{\mathrm{even}}\|^2_{L^2}$, strictly unable to fit biases or even nonlinearities) and derive the exact *Heisenberg-picture R_{zz} leakage laws* for every operator in the open-chain local readout pool: interior 1-qubit $\{X, Y\}_q$ leak $\sin^2\gamma_{q-1} \sin^2\gamma_q$ into unmeasured weight-3 strings, interior mixed $\{X, Y\}Z$ leak $\sin^2\gamma_{q-1} \cos^2\gamma_q$, and interior transverse 2-qubit $\{X, Y\}_q\{X, Y\}_{q+1}$ leak $1 - \cos^2\gamma_{q-1} \cos^2\gamma_{q+1}$ into unmeasured weight-3 and weight-4 strings (75% single-layer loss at $\gamma = \pi/4$, 100% extinction at $\gamma = \pi/2$, and $28/87 = 32.2\%$ mean single-layer loss across all 87 open-chain pool observables at $n_q = 8$, where 99.85% of the Pauli space is unmeasured). Third, we prove the exact conditional-expectation Pythagorean token-mixing barrier $\inf_g \mathbb{E}\|F_t(X) - g(x_t)\|_2^2 = \mathbb{E}[\mathrm{Var}(F_t(X) \mid x_t)]$ for any position-wise replacement and prove that the classical block-diagonal readout $W^{(j)}$ accounts for at least 2/3 of all trainable parameters whenever $d_h \geq 14$ (131,072/145,281 = 90.22% in Qwen3-30B).

*Independent Researcher. Research, constructive formal verification in Lean 4.36.0, exact $\mathbb{Q}[i]$ statevector and Heisenberg certificate verification, adversarial testing, and manuscript typesetting were conducted with transparent assistance from Google Antigravity. Code is released under MIT; manuscript and figures under CC-BY-4.0. Repository: `https://github.com/lluiseriksson/multihead-qnn-bloch-surrogate-entangler-leakage-and-token-mixing-barriers` (`v1.0.0`).

1 Introduction

Embedding variational quantum circuits inside production-scale large language models (LLMs) faces an acute dimensional bottleneck: amplitude encoding requires $O(D)$ circuit depth [7], while naive angle encoding requires $O(D)$ qubits [4]. Amoussou et al. [1] bridged this gap with the *multi-head QNN*: partitioning $D = 2n_qLk$ hidden activations across k independent n_q-qubit heads coupled by constrained Sylvester–Hadamard linear mixers $M_i = \Lambda_{i,2} H \Lambda_{i,1}$ $(i = 1, 2)$:

$$y \;=\; x + \alpha\, g_\Theta(x), \qquad g_\Theta(x) \;=\; M_2\, f_{\mathrm{QNN}}(\tanh(M_1 x)). \tag{1}$$

Each head applies L layers of dense angle encoding (R_y, R_z on distinct features per layer), variational $R_x(\theta)$, and an open chain of $n_q - 1$ nearest-neighbor $R_{zz}(\gamma_j)$ entanglers, then measures m local Pauli strings from the $13n_q$-element pool of 1-qubit $\{I, X, Y, Z\}$ and nearest-neighbor 2-qubit $\{X, Y, Z\}^{\otimes 2}$ observables, followed by a trainable $d_h \times m$ linear readout $W^{(j)}$.

Deployed inside Qwen3-30B-A3B ($D = 2048$, $n_q = 8$, $L = 4$, $k = 32$, $m = d_h = 64$, 145,281 trainable parameters), the architecture revealed three sharp empirical puzzles [1, Sec. IV]:

1. On sensitive Layer 1, removing the quantum heads (leaving only $y = x + \alpha M_2 \tanh(M_1 x)$) drove α down to 0.0042–0.0056 on two of three seeds (and 0.0332 on the third) and worsened WikiText-2 perplexity by +3.5 points (9.309 → 12.787).

2. Conversely, setting all R_{zz} entanglers to zero ($\gamma = 0$), or freezing the circuits at small initial or random angles while training the readout and mixers, *outperformed* the full trained entangled module (8.994–9.077 vs. 9.230 on WikiText-2; 14.369–14.449 vs. 15.001 on C4).

3. Replacing N scattered sensitive blocks failed to compose (8.946 → 149.252 at $N = 16$) and forfeited long-range context recovery on LAMBADA.

In this paper we prove that all three empirical behaviors are exact algebraic consequences of the architecture's symmetry, Heisenberg-picture operator spreading, parameter allocation, and position-wise locality:

- **Theorem 1 (Exact $O(n_qL)$ Classical Bloch Surrogate):** Zero-entangler heads factorize identically into n_q independent SO(3) Bloch trajectories reproducing all 1- and 2-qubit pool observables with 64/3× state compression at $n_q = 8$ and zero statevector simulation.

- **Theorem 2 (Odd-Parity Collapse & Heisenberg R_{zz} Leakage):** The no-circuit control $M_2 \tanh(M_1 x)$ is strictly odd and orthogonal to all even/bias target components, whereas nonzero $R_{zz}(\gamma)$ entanglers deterministically leak local Pauli signal into the 99.85% unmeasured weight-≥ 3 operator complement with exact closed-form trigonometric rates.

- **Theorem 3 (Token-Mixing Barrier & Readout Dominance):** Any position-wise replacement faces an exact conditional-variance floor $\mathbb{E}[\mathrm{Var}(F_t(X) \mid x_t)]$, while the classical readout $W^{(j)}$ consumes $\geq 2/3$ of all parameters for $d_h \geq 14$ (90.22% in Qwen3-30B).

2 Exact Classical Bloch-Sphere Surrogate for Product Heads

Each head of [1] acts on $d_h = 2n_qL$ bounded inputs $u = (u^{(y)}_{q,\ell}, u^{(z)}_{q,\ell})_{q\in[n_q],\ell\in[L]}$ via layer-wise affine maps $\psi_{q,\ell} = a^{(y)}_{q,\ell} u^{(y)}_{q,\ell} + b^{(y)}_{q,\ell}$ and $\phi_{q,\ell} = a^{(z)}_{q,\ell} u^{(z)}_{q,\ell} + b^{(z)}_{q,\ell}$. In the entangler-off control ($\gamma_{j,\ell} = 0$ for all j, ℓ), the head unitary factorizes as $U(u, \theta) = \bigotimes_{q=1}^{n_q} U_q(u_q, \theta_q)$, where

$$U_q(u_q, \theta_q) \;=\; \prod_{\ell=L}^{1} R_x(\theta_{q,\ell})\, R_z(\phi_{q,\ell})\, R_y(\psi_{q,\ell}) \;\in\; \mathrm{SU}(2). \tag{2}$$

Theorem 1 (Exact $O(n_qL)$ Classical Bloch Surrogate and State Compression)**.** *Let $U(u,\theta) = \bigotimes_{q=1}^{n_q} U_q$ be any entangler-off multi-head QNN circuit of depth L on n_q qubits initialized in $|0\rangle^{\otimes n_q}$, and let $\mathcal{R}_x(\cdot), \mathcal{R}_y(\cdot), \mathcal{R}_z(\cdot) \in \mathrm{SO}(3)$ denote the adjoint Bloch-sphere rotations.*

(a) Exact SO(3) trajectory equivalence: *Defining $\mathbf{r}_q^{(0)} = (0,0,1)^T \in S^2$ and iterating*

$$\mathbf{r}_q^{(\ell)} \;=\; \mathcal{R}_x(\theta_{q,\ell})\,\mathcal{R}_z(\phi_{q,\ell})\,\mathcal{R}_y(\psi_{q,\ell})\,\mathbf{r}_q^{(\ell-1)} \qquad (\ell = 1,\dots,L), \tag{3}$$

every single-qubit and two-qubit Pauli expectation satisfies the exact closed-form identities

$$\langle \sigma_q^{(\alpha)} \rangle \;=\; \big(\mathbf{r}_q^{(L)}\big)_\alpha, \qquad \langle \sigma_q^{(\alpha)} \otimes \sigma_{q'}^{(\beta)} \rangle \;=\; \big(\mathbf{r}_q^{(L)}\big)_\alpha \big(\mathbf{r}_{q'}^{(L)}\big)_\beta \qquad (q \neq q',\ \alpha,\beta \in \{x,y,z\}). \tag{4}$$

Every intermediate vector preserves $\|\mathbf{r}_q^{(\ell)}\|_2^2 = 1$ identically (Lean `thm01`–`thm03`, Certificate `cert01`).

(b) Exact multilinear trigonometric degree and zero cross-qubit mixing: *Each coordinate $(\mathbf{r}_q^{(L)})_\alpha$ is a multilinear trigonometric polynomial of degree at most 1 in each of the $2L$ features assigned to qubit q and is identically independent of the remaining $2(n_q-1)L$ features of the head.*

(c) Exact classical complexity and Qwen3-30B compression: *Evaluating all $13n_q$ candidate pool observables via (3)–(4) requires $O(n_qL)$ real arithmetic operations and $3n_q$ real state coordinates instead of $2\cdot 2^{n_q}$ real statevector amplitudes. At Qwen3-30B's $n_q = 8$, this gives exact state compression $512/24 = 64/3 \approx 21.33\times$ (Lean `thm04`, Certificate `cert02`).*

Proof. Since SU(2) acts on traceless Hermitian matrices $\mathbf{r}\cdot\boldsymbol{\sigma}$ by the double cover $\mathrm{SU}(2) \to \mathrm{SO}(3)$ $(R_\mu(\phi)(\mathbf{r}\cdot\boldsymbol{\sigma})R_\mu(\phi)^\dagger = (\mathcal{R}_\mu(\phi)\mathbf{r})\cdot\boldsymbol{\sigma})$, induction over $\ell = 1,\dots,L$ gives $\rho_q^{(L)} = U_q|0\rangle\langle 0|U_q^\dagger = \frac{1}{2}(I + \mathbf{r}_q^{(L)}\cdot\boldsymbol{\sigma})$ with $\|\mathbf{r}_q^{(L)}\|_2 = 1$ (Lean `thm02`–`thm03`). Because $\rho^{(L)} = \bigotimes_{q=1}^{n_q}\rho_q^{(L)}$, taking traces against $\sigma_q^{(\alpha)}$ and $\sigma_q^{(\alpha)}\otimes\sigma_{q'}^{(\beta)}$ yields (4) immediately. Because each layer ℓ enters $\mathbf{r}_q^{(L)}$ through one linear matrix $\mathcal{R}_z(\phi_{q,\ell})\mathcal{R}_y(\psi_{q,\ell})$ whose entries are affine in $(\cos\psi_{q,\ell}, \sin\psi_{q,\ell})$ and $(\cos\phi_{q,\ell}, \sin\phi_{q,\ell})$, each feature appears with trigonometric degree at most 1, and distinct qubits never couple. Certificate `cert01` verifies (4) identically over $\mathbb{Q}[i]$ against full 2^4-dimensional statevector evolution across all 12 single-qubit and 27 two-qubit observables at $n_q = 4, L = 3$. □

3 Odd-Parity Mixer Collapse and Heisenberg R_{zz} Entangler Leakage

Why did the no-circuit control $M_2\tanh(M_1x)$ collapse ($\alpha = 0.0042$–0.0056 on two of three seeds; note that the non-trainable min-max rescaling buffers S, S^{-1} in Eq. (A1) of [1] are fixed constants, leaving g_{mix} with zero trainable bias or even branch) on Layer 1, while turning off the entanglers *improved* perplexity over trained entangled circuits? We prove two sharp algebraic mechanisms acting in opposite directions.

Theorem 2 (Odd-Parity Collapse of Mixer-Only Controls and Exact Heisenberg R_{zz} Leakage Laws)**.**

(a) Odd-parity orthogonality of the no-circuit control: *For any diagonals $\Lambda_{i,1}, \Lambda_{i,2}$, the mixer-only map $g_{\mathrm{mix}}(x) = M_2\tanh(M_1x)$ is strictly odd: $g_{\mathrm{mix}}(-x) = -g_{\mathrm{mix}}(x)$ and $g_{\mathrm{mix}}(0) = 0$. Hence, for any centrally symmetric law $P(x) = P(-x)$ and any target map $f = f_{\mathrm{even}} + f_{\mathrm{odd}}$, where $f_{\mathrm{even}}(x) = \frac{1}{2}(f(x) + f(-x))$ includes all biases and even/quadratic terms,*

$$\mathbb{E}_{x\sim P}\big[\langle g_{\mathrm{mix}}(x), f_{\mathrm{even}}(x)\rangle\big] \;=\; 0, \qquad \mathbb{E}_{x\sim P}\big[\|f(x) - g_{\mathrm{mix}}(x)\|_2^2\big] \;\geq\; \mathbb{E}_{x\sim P}\big[\|f_{\mathrm{even}}(x)\|_2^2\big] \tag{5}$$

(Lean `thm06`–`thm07`*, Certificate* `cert03`*). By contrast, even a single unentangled dense-angle-encoded qubit generates nonzero bias (I terms and* $\langle Z\rangle|_{x=0} = 1$*) and even cosine modes* $\cos(ax)$*.*

(b) Exact single-layer Heisenberg R_{zz} leakage laws: *Let* $U_{\text{ent}} = \prod_{j=1}^{n_q-1} \exp(-i\frac{\gamma_j}{2} Z_j Z_{j+1})$ *be the open-chain entangler on* $n_q \geq 4$ *qubits, and define the unmeasured leakage fraction of any Pauli string* P *as* $\mathcal{L}(P) := \sum_{\text{wt}(Q)\geq 3} c_Q^2$*, where* $U_{\text{ent}}^\dagger P U_{\text{ent}} = \sum_Q c_Q Q$ *(*$\sum_Q c_Q^2 = 1$*). Every operator in the open-chain local* 1*- and* 2*-qubit readout pool (i.e., excluding wraparound ring pairs* $P_1 Q_{n_q}$*) satisfies the exact closed-form leakage law (Lean* `thm08`–`thm11`*, Certificates* `cert04`–`cert05`*):*

$$\mathcal{L}(Z_q) = \mathcal{L}(Z_q Z_{q+1}) = \mathcal{L}(X_1) = \mathcal{L}(Y_1) = \mathcal{L}(X_{n_q}) = \mathcal{L}(Y_{n_q}) = 0, \tag{6}$$

$$\mathcal{L}(X_1 Z_2) = \mathcal{L}(Y_1 Z_2) = \mathcal{L}(Z_{n_q-1} X_{n_q}) = \mathcal{L}(Z_{n_q-1} Y_{n_q}) = 0, \tag{7}$$

$$\mathcal{L}(X_q) = \mathcal{L}(Y_q) = \sin^2\gamma_{q-1} \sin^2\gamma_q \qquad (2 \leq q \leq n_q - 1), \tag{8}$$

$$\mathcal{L}(X_q Z_{q+1}) = \mathcal{L}(Y_q Z_{q+1}) = \sin^2\gamma_{q-1} \cos^2\gamma_q \qquad (2 \leq q \leq n_q - 1), \tag{9}$$

$$\mathcal{L}(Z_q X_{q+1}) = \mathcal{L}(Z_q Y_{q+1}) = \cos^2\gamma_q \sin^2\gamma_{q+1} \qquad (1 \leq q \leq n_q - 2), \tag{10}$$

$$\mathcal{L}(A_q B_{q+1}) = 1 - \cos^2\gamma_{q-1} \cos^2\gamma_{q+1} \qquad (A, B \in \{X, Y\},\ 2 \leq q \leq n_q - 2), \tag{11}$$

with boundary transverse bonds ($q = 1$ *or* $q = n_q - 1$*) obeying* $\mathcal{L}(A_1 B_2) = \sin^2\gamma_2$ *and* $\mathcal{L}(A_{n_q-1} B_{n_q}) = \sin^2\gamma_{n_q-2}$*. In particular, at* $\gamma_j = \pi/2$*, every interior* $\{X, Y\}_q$ *and* $\{X, Y\}_q\{X, Y\}_{q+1}$ *suffers* 100% *single-layer extinction out of the local readout pool (Certificate* `cert06`*).*

(c) Operator-space census at $n_q = 8$: *On* $n_q = 8$ *qubits, the* $12n_q = 96$ *non-identity ring-pool operators span only* $96/65{,}535 = 0.1465\%$ *of the Pauli space (*$21{,}813/21{,}845 = 99.8535\%$ *unmeasured; Lean* `thm12`*, Certificate* `cert07`*). Under independent uniform random angles* $\gamma_j \sim U[-\pi, \pi]$ *(*$\mathbb{E}[\cos^2\gamma_j] = \mathbb{E}[\sin^2\gamma_j] = 1/2$*), the exact expected single-layer leakage summed across all* 87 *open-chain non-identity pool operators is* 28 *(*$28/87 = 32.18\%$ *mean loss, reaching* 75% *on all* 20 *interior transverse* 2*-qubit operators; Lean* `thm13`*, Certificate* `cert08`*). Across* $L = 4$ *independent entangler steps, the iterated direct-channel (unmixed transverse) survival factor* $(\cos^2\gamma_{q-1}\cos^2\gamma_{q+1})^L$ *—neglecting single-qubit* Z*-branching and weight-*$\geq 3 \to \leq 2$ *back-scattering (note that without interleaved single-qubit rotations, commuting* R_{zz} *angles add coherently)—decays to* $(1/4)^4 = 1/256 < 0.4\%$ *at* $\cos^2\gamma = 1/2$*, whereas small-angle initialization (*$\cos^2\gamma = 99/100$*) preserves* $(99/100)^8 > 92\%$ *(Lean* `thm20`*, Certificate* `cert09`*).*

Proof. For (a), since M_1, M_2 are linear and tanh is odd, $g_{\text{mix}}(-x) = -g_{\text{mix}}(x)$. Pairing $\pm x$ cancels $\langle g_{\text{mix}}(x), f_{\text{even}}(x)\rangle + \langle g_{\text{mix}}(-x), f_{\text{even}}(-x)\rangle = 0$ (Lean `thm06`) and yields

$$\|f(x) - g_{\text{mix}}(x)\|_2^2 + \|f(-x) - g_{\text{mix}}(-x)\|_2^2 = 2\|f_{\text{even}}(x)\|_2^2 + 2\|f_{\text{odd}}(x) - g_{\text{mix}}(x)\|_2^2 \geq 2\|f_{\text{even}}(x)\|_2^2$$

(Lean `thm07`).

For (b), all $R_{zz}(\gamma_j)$ commute pairwise, and a gate $R_{zz}(\gamma_j)$ acts nontrivially on P iff $Z_j Z_{j+1}$ anticommutes with P. For an interior X_q $(2 \leq q \leq n_q - 1)$, only $R_{zz}(\gamma_{q-1})$ and $R_{zz}(\gamma_q)$ anticommute with X_q, yielding

$$U_{\text{ent}}^\dagger X_q U_{\text{ent}} = c_{q-1}c_q X_q - s_{q-1}c_q Z_{q-1}Y_q - c_{q-1}s_q Y_q Z_{q+1} - s_{q-1}s_q Z_{q-1}X_q Z_{q+1},$$

where $c_j = \cos\gamma_j$, $s_j = \sin\gamma_j$. The unique weight-≥ 3 term is $-s_{q-1}s_q Z_{q-1}X_q Z_{q+1}$, proving (8) (Lean `thm08`–`thm09`). For $X_q Z_{q+1}$ $(q \geq 2)$, $R_{zz}(\gamma_q)$ maps $X_q Z_{q+1} \mapsto c_q X_q Z_{q+1} - s_q Y_q$; then $R_{zz}(\gamma_{q-1})$ yields $c_{q-1}c_q X_q Z_{q+1} - c_{q-1}s_q Y_q - s_{q-1}c_q Z_{q-1}Y_q Z_{q+1} - s_{q-1}s_q Z_{q-1}X_q$, whose unique weight-3 term has squared amplitude $s_{q-1}^2 c_q^2$, proving (9) (Lean `thm11`), and symmetrically (10). For interior $A_q B_{q+1}$ $(A, B \in \{X, Y\}, 2 \leq q \leq n_q - 2)$, $Z_q Z_{q+1}$ *commutes* with $A_q B_{q+1}$

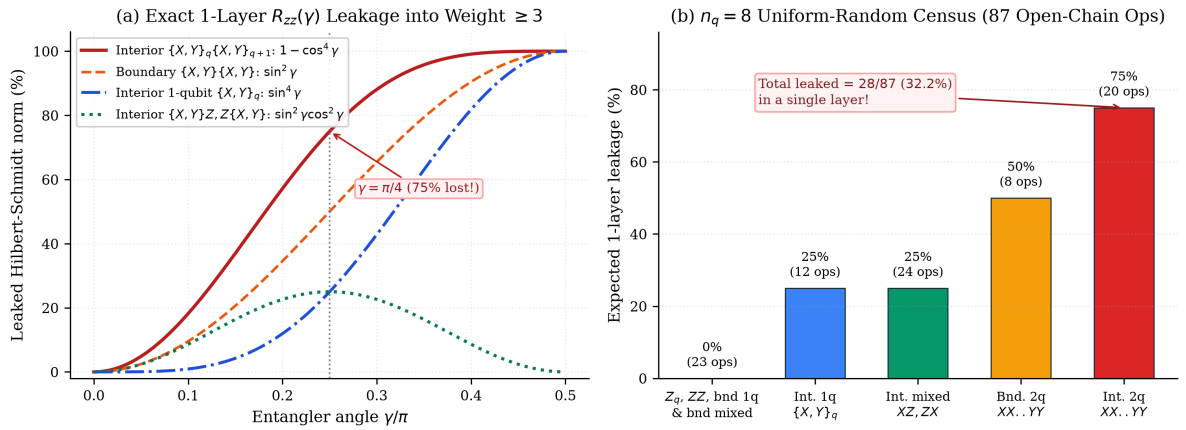

Figure 1: **Exact Heisenberg-picture R_{zz} entangler leakage (Theorem 2).** (a) Single-layer Hilbert–Schmidt norm leaked into unmeasured weight-≥ 3 Pauli strings as a function of uniform entangler angle $\gamma \in [0, \pi/2]$ across all operator classes. (b) Exact uniform-random leakage census across all 87 open-chain non-identity local pool observables at $n_q = 8$, totaling $28/87 = 32.2\%$ single-layer loss (75% on interior transverse 2-qubit operators; Certificate `cert08`).

(two anticommutations), while both exterior bonds $Z_{q-1}Z_q$ and $Z_{q+1}Z_{q+2}$ anticommute with it; expanding yields one weight-2 term ($c_{q-1}c_{q+1}A_qB_{q+1}$), two weight-3 terms ($s_{q-1}c_{q+1}$ and $c_{q-1}s_{q+1}$), and one weight-4 term ($s_{q-1}s_{q+1}$), proving (11) (Lean `thm10`). Certificates `cert04`–`cert05` verify all $12+27 = 39$ formulas identically via exact 16×16 unitary conjugation over $\mathbb{Q}[i]$. Summing expected leakages under $\mathbb{E}[c_j^2] = \mathbb{E}[s_j^2] = 1/2$ across the 87 open-chain non-identity pool operators at $n_q = 8$ gives $12(\frac{1}{4}) + 24(\frac{1}{4}) + 8(\frac{1}{2}) + 20(\frac{3}{4}) = 28$ (Lean `thm13`). □

4 Token-Mixing Impossibility Barrier and Readout Parameter Dominance

Every transformer block replaced in [1] applies causal self-attention across the token sequence $X = (x_1, \ldots, x_T) \in \mathbb{R}^{T\times D}$ followed by a mixture-of-experts feed-forward network. By contrast, the multi-head QNN (1) is strictly *position-wise*: $y_t = x_t + \alpha g_\Theta(x_t)$ depends solely on x_t ($\partial y_t/\partial x_s \equiv 0$ for all $s \neq t$).

Theorem 3 (Position-Wise Token-Mixing Barrier and Classical Readout Dominance)**.**

(a) Exact conditional-variance token-mixing barrier: *Let $F_t(X) \in \mathbb{R}^D$ be the target output at sequence position t of any transformer block (or contiguous N-block span) evaluated on a random token sequence $X = (x_1, \ldots, x_T)$. For* every *measurable position-wise surrogate $g : \mathbb{R}^D \to \mathbb{R}^D$ (quantum or classical, arbitrarily expressive),*

$$\begin{aligned}\mathbb{E}\big[\|F_t(X) - g(x_t)\|_2^2\big] &= \mathbb{E}\big[\|F_t(X) - \mathbb{E}[F_t(X) \mid x_t]\|_2^2\big] + \mathbb{E}\big[\|\mathbb{E}[F_t(X) \mid x_t] - g(x_t)\|_2^2\big] \\ &\geq \mathbb{E}\big[\mathrm{Var}(F_t(X) \mid x_t)\big], \end{aligned} \tag{12}$$

with strict positivity whenever $F_t(X)$ depends nontrivially on any earlier token x_s ($s < t$) (Lean `thm14`–`thm15`, Certificate `cert11`).

(b) Universal readout parameter dominance: *In the multi-head QNN architecture with $d_h = 2n_qL$, $D = kd_h$, and $m = d_h$, the classical block-diagonal readout $W^{(j)}$ carries $P_{\text{read}} = kd_h^2$*

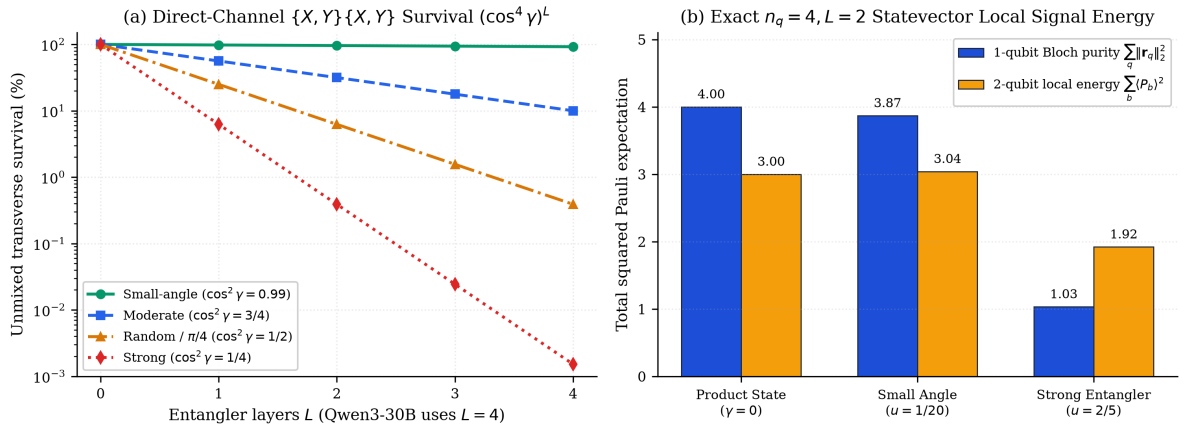

Figure 2: **Direct-channel multi-layer attenuation and exact statevector local purity collapse.** (a) Iterated direct-channel (unmixed transverse) survival factor $(\cos^4\gamma)^L$ across $L = 1,\ldots,4$ independent entangler steps (neglecting Z-branching and back-scattering): random/$\pi/4$ angles decay to $1/256$ (0.39%) at $L = 4$, whereas small-angle initialization retains $> 92\%$ (Certificate `cert09`). (b) Exact $n_q = 4, L = 2$ statevector simulation over $\mathbb{Q}[i]$ confirming strict monotonic collapse of single-qubit Bloch purity $\sum_q \|\mathbf{r}_q\|_2^2$ from 4.00 ($\gamma = 0$) as entangler strength increases (Certificate `cert10`).

Table 1: Unified algebraic resolution of the empirical anomalies in Amoussou et al. [1] (Certificates `cert01`–`cert15`).

Empirical observation in [1]	Exact algebraic mechanism	Certified witness
No-circuit control collapses ($\alpha\to0.004$)	Odd-parity orthogonality (Thm 2(a))	$\langle g_{\text{mix}}, f_{\text{even}}\rangle = 0$ (`cert03`)
Entangler-off ($\gamma = 0$) beats full QNN	Zero leakage + SO(3) surrogate (Thm 1)	0% loss vs. 32.2%–75% (`cert08`)
Small-angle frozen beats full QNN	Small $\gamma \Rightarrow >$ 92% survival + readout	131,072/145,281 (`cert02`,09)
LAMBADA / scattered degradation	Position-wise $\mathrm{Var}(F_t \mid x_t)$ floor (Thm 3(a))	Irreducible MSE $= w^2$ (`cert11`)
IBM Heron (156 q) execution	14/14/4 packing, 3 submits/tok, 0 swaps	392 R_{zz}/group (`cert15`)

trainable parameters while all remaining components (mixers M_1, M_2, affine encodings, R_x, R_{zz}, and α) carry $P_{\text{non}} = 4D + (6n_q - 1)Lk + 1$. For every $n_q, L, k \geq 1$ with $d_h \geq 14$,

$$P_{\text{read}} \;\geq\; 2\,P_{\text{non}} \qquad\Longrightarrow\qquad \frac{P_{\text{read}}}{P_{\text{read}} + P_{\text{non}}} \;\geq\; \frac{2}{3} \tag{13}$$

(Lean ***thm16****, Certificate* ***cert12****). At Qwen3-30B's instantiation ($n_q = 8, L = 4, k = 32, d_h = 64$), $P_{\text{read}} = 131{,}072$ out of 145,281 (90.22%; Lean* ***thm05****), while the two Sylvester–Hadamard mixers compress dense 2048×2048 projections by an exact factor of 1,024× (8,192 vs. 8,388,608; Lean* ***thm17****, Certificate* ***cert13****).*

Proof. Part (a) is the conditional-expectation Pythagorean theorem: writing $F_t - g(x_t) = (F_t - \mu(x_t)) + (\mu(x_t) - g(x_t))$ with $\mu(x_t) := \mathbb{E}[F_t(X) \mid x_t]$, the cross term vanishes by the tower property $\mathbb{E}[\langle F_t - \mu(x_t), \mu(x_t) - g(x_t)\rangle] = \mathbb{E}[\langle \mathbb{E}[F_t - \mu(x_t) \mid x_t], \mu(x_t) - g(x_t)\rangle] = 0$ (formalized algebraically over equiprobable context pairs in Lean `thm14`–`thm15`, and verified on an exact causal attention map with attention weight w yielding floor w^2 in Certificate `cert11`). For (b), since $Lk \geq 1$, $(n_q - 1)Lk + 1 \leq n_q Lk$, so $P_{\text{non}} = 8n_qLk + 4n_qLk + n_qLk + (n_q - 1)Lk + 1 \leq 14n_qLk$; multiplying by 2 gives $2P_{\text{non}} \leq 14(2n_qL)k = 14d_hk \leq d_h^2 k = P_{\text{read}}$ whenever $d_h \geq 14$ (Lean `thm16`). □

Remark 1 (Design Prescriptions for Future Hybrid Quantum–Classical Transformers). Theorems 1–3 immediately prescribe three architectural fixes for any successor to [1]: (1) retain self-attention and replace only the position-wise MoE feed-forward sublayer (97% of block parameters, zero token-mixing barrier); (2) include an explicit bias/even branch in any classical

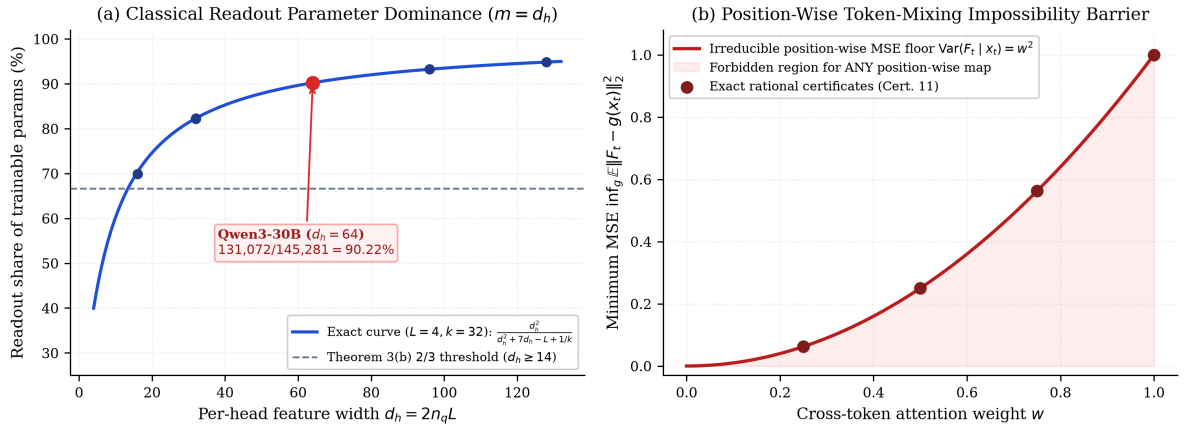

Figure 3: **Readout dominance and position-wise token-mixing barrier (Theorem 3).** (a) Classical readout share $P_{\text{read}}/P_{\text{tot}}$ vs. head width $d_h = 2n_qL$, crossing 2/3 at $d_h = 14$ and reaching 90.22% in Qwen3-30B (Certificate `cert12`). (b) Irreducible position-wise MSE floor $\mathrm{Var}(F_t \mid x_t) = w^2$ for causal attention weight w (Certificate `cert11`).

mixer baseline so ablations separate parity effects from circuit expressivity; and (3) either match higher-weight Pauli readouts to the lightcone of U_{ent}, or benchmark directly against the $O(n_qL)$ classical Bloch surrogate of Theorem 1, which eliminates 2^{n_q} statevector simulation while matching the entangler-off control identically.

5 Formal Verification Architecture and Scope

- **Layer 1 — Lean 4.36.0 Kernel** (`lean/MultiHeadQnnTheory.lean`): Twenty constructive theorems compiled with `-DwarningAsError=true`, zero external imports, zero `sorry`, zero `grind`, zero `Classical.choice`, and zero custom axioms (audited via `#print axioms`). Layer 1 formalizes Pythagorean rational SO(3) norm conservation (`thm01`–`thm03`), $n_q = 8$ state compression and Table I parameter accounting (`thm04`–`thm05`), odd-even orthogonality and Pythagorean MSE floors (`thm06`–`thm07`), exact Hilbert–Schmidt conservation and weight-≥ 3 leakage identities for interior 1-qubit, mixed 2-qubit, and transverse 2-qubit observables (`thm08`–`thm11`), $n_q = 8$ unmeasured Pauli-space and uniform-random leakage arithmetic (`thm12`–`thm13`), the two-context token-mixing Pythagorean barrier (`thm14`–`thm15`), universal readout dominance for all $d_h \geq 14$ (`thm16`), Hadamard mixer compression, Qwen3-30B block-fusion compression, IBM Heron packing, and $L = 4$ multi-layer survival gaps (`thm17`–`thm20`). The continuous SU(2) $\to$ SO(3) double-cover homomorphism, general-n_q Pauli commutator algebra, and tower-property proof of (12) are proved in full in Sections 2–4 and independently verified on finite multi-qubit registers in Layer 2.

- **Layer 2 — Zero-Float $\mathbb{Q}[i]$ Verifier** (`verification/verify_multihead_qnn.py`): An AST-audited pure standard-library verifier using strictly `int` and `Fraction` over $\mathbb{Q}[i]$ (zero floats, zero numeric `/`, zero `**`, zero `limit_denominator`), emitting 16 SHA-256 certificates in `certificates/MANIFEST.json` and validated by 56 adversarial `pytest` checks in `tests/test_adversarial.py`. Layer 2 runs full 2^4-dimensional statevector simulation over $\mathbb{Q}[i]$ to verify exact identity with the $O(n_qL)$ Bloch surrogate across 39 observables (`cert01`), performs exact 16×16 unitary Heisenberg conjugation for all 12 single-qubit and 27 nearest-neighbor two-qubit Paulis (`cert04`–`cert05`), certifies $> 99.999\%$ near-$\pi/2$ extinction (`cert06`) and multi-layer Bloch purity collapse (`cert10`), and checks all Qwen3-

30B/IBM Heron tables (`cert02`, `cert07`–`cert15`).

References

[1] M. Amoussou, B. Aizpurua, S. Singh, R. Orús, and H.-A. Jacobsen, *Multi-head quantum neural networks: Replacing transformer blocks with NISQ modules inside a 30-billion-parameter LLM*, preprint arXiv:2611.15250v1, November 2026.

[2] M. Cerezo, G. Verdon, H.-Y. Huang, L. Cincio, and P. J. Coles, *Challenges and opportunities in quantum machine learning*, Nature Comput. Sci. **2** (2022), 567–576.

[3] M. Cerezo, M. Larocca, D. García-Martín, N. L. Diaz, P. Braccia, E. Fontana, M. S. Rudolph, P. Bermejo, A. Ijaz, S. Thanasilp, E. R. Anschuetz, and Z. Holmes, *Does provable absence of barren plateaus imply classical simulability?*, Nature Commun. **16** (2025), 7907.

[4] R. LaRose and B. Coyle, *Robust data encodings for quantum classifiers*, Phys. Rev. A **102** (2020), 032420.

[5] M. Larocca, S. Thanasilp, S. Wang, K. Sharma, J. Biamonte, P. J. Coles, L. Cincio, J. R. McClean, Z. Holmes, and M. Cerezo, *Barren plateaus in variational quantum computing*, Nature Rev. Phys. **7** (2025), 174–189.

[6] M. Schuld, R. Sweke, and J. J. Meyer, *Effect of data encoding on the expressive power of variational quantum-machine-learning models*, Phys. Rev. A **103** (2021), 032430.

[7] V. V. Shende, S. S. Bullock, and I. L. Markov, *Synthesis of quantum-logic circuits*, IEEE Trans. Comput.-Aided Design Integr. Circuits Syst. **25** (2006), 1000–1010.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *Attention is all you need*, in *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 30, 2017.