HIGHER-ORDER TOKEN INTERACTIONS VIA QUANTUM ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A single self-attention layer scores token *pairs* and aggregates them linearly over values, and known lower bounds put specific order-k tasks out of reach for one such layer at budget $mHp = o(N/\log\log N)$. We introduce **Quantum Higher-Order Attention (QHA)**, a shallow, hardware-realizable quantum attention head that — via data re-uploading and an all-to-all non-Clifford entangler — synthesizes order-k token interactions *inside* the circuit and exposes them on a local single-qubit read-out. We prove two *complementary* results, deliberately not billed as a separation theorem because they concern different targets: no budget-limited attention layer computes the order-k *matching* family, while a circuit in the QHA gate set realizes every order-k *monomial* at depth $O(\log k)$; and a barren-plateau-free trainability guarantee for the shallow local-design variant — the all-to-all variant we benchmark does show exponentially decaying gradients. Empirically we tune the classical head's learning rate, width and training length to maximize *its own* accuracy, an advantage we do not grant QHA, and find the gap is *graded in* k: a tuned attention layer learns order-3 parity (0.93), so no separation holds there; from order 4 QHA (296 parameters) leads $1.00/1.00/0.92$ against $0.74/0.58/0.49$ at $k{=}4/5/6$; and at order 6 every attention-paradigm model we test — explicit high-order attention and a $16\times$ larger two-layer Transformer included — sits at chance. Non-attention models do solve it (a degree-matched HOFM at $2.3\times$ QHA's parameters reaches 1.00), so this is a separation *within the attention paradigm*, at the smallest parameter budget of any model that gets there. A trained order-3 head runs on IBM Heron at $\geq 95\%$ accuracy through 14 qubits under a disclosed shot budget, and QHA is the most accurate and most compact detector on genetic epistasis and learning-parity-with-noise; on graph triangle detection, once a protocol error is corrected, no model separates and we report that as a limitation. Our claim is an *inductive-bias advantage* and an *empirical learning separation* against classical attention, not an exponential speedup.

1 INTRODUCTION

The self-attention layer at the heart of the Transformer (Vaswani et al., 2017) scores a pair of tokens by a dot product $q_i^\top k_j$ and mixes their values accordingly. Its *scoring primitive is therefore pairwise*, and the layer aggregates those scores linearly over values. The layer as a whole is not a degree-2 polynomial — the softmax is non-linear and couples all tokens — but the pairwise primitive is what the known lower bounds exploit: on the specific tasks for which such bounds exist, a single budget-limited layer provably fails, and the constructions that do reach order k pay in depth or explicit $O(n^k)$ cost. (No result we know of shows that *every* order-k function forces $O(n^k)$.) Sanford et al. (2023) make this precise — a single softmax attention layer needs width/heads/precision polynomial in the sequence length to compute even a single order-3 matching predicate — and Kozachinskiy et al. (2025) show that standard attention provably cannot solve order-3 compositional tasks without super-quadratic resources. Yet many problems are intrinsically high-order: k-ary matching and function composition (Sanford et al., 2023; Kozachinskiy et al., 2025), parity-like correlations that underlie algorithmic reasoning (Hahn, 2020; Bhattamishra et al., 2023), and k-body interactions in physics, chemistry, and biology.

Quantum circuits are a natural home for high-order interactions: a parity over a k-subset, the canonical order-k Boolean function, is computed by a constant-depth circuit, and the reachable polynomial

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

degree of a data-re-uploading circuit grows with the number of re-uploads (Schuld et al., 2021). This raises a concrete question that, surprisingly, the quantum-machine-learning literature has not answered: *can a shallow quantum attention head represent and learn order-k token interactions that a single classical attention layer of equal or larger budget provably cannot, and does the advantage survive on real hardware?*

Existing quantum Transformers do not settle this (Sec. 2). None proves an expressivity separation against classical attention, most are evaluated on small data without parameter-matched baselines, and a systematic study from the same community concludes that bolting a variational circuit into attention yields "only marginal gains ... while significantly increasing parameter count" (Chen & Kuo, 2026). The bar set by the field's own critics (Bowles et al., 2024) — matched budgets, honest baselines, disclosed shots, and a clear statement of what is and is not classically simulable — is rarely met.

We meet it. We introduce **Quantum Higher-Order Attention (QHA)**, a shallow quantum attention head whose single-qubit read-out realizes order-k token interactions inside the circuit, and we evaluate it under a strict matched-budget, disjoint-split protocol with a real-hardware demonstration.

Contributions.

- **Mechanism.** QHA, a shallow quantum attention head whose *local* read-out realizes order-k token interactions inside the circuit (Sec. 3).
- **Theorem 1 (complementary order-k representation results).** No single classical self-attention layer with $mHp = o(N/\log\log N)$ computes the order-k matching family $\textsc{Match}_k$, $k \geq 3$; and a circuit in the QHA gate set realizes every order-k monomial on a single-qubit read-out at depth $O(\log k)$. The two targets differ, so this is *not* a proved separation (Sec. 4).
- **Theorem 2 (trainability).** The local-design instantiation (local read-out, $O(\log n)$ depth) is barren-plateau-free; we are explicit that the all-to-all instantiation we benchmark is trained empirically (Secs. 4, 5.1).
- **Evidence.** A learnability gap against a *tuned* attention baseline (best of 20 configurations, up to $2\times$ QHA's training budget) that is graded in k: absent at k=3, significant from k=4, and total at k=6, where every attention-paradigm model tested is at chance; plus IBM hardware validation and an honest treatment of dequantization (Secs. 5, 7).

2 Related Work

Quantum Transformers and quantum attention. The survey of Zhang et al. (2025) splits the field into PQC-based and QLA-based paradigms. PQC-based models put a circuit on the $Q/K/V$ projections or the scoring function while keeping softmax classical (QSANN, QMSAN, QASA, QPSAN, QClusformer), replace the normalization operator (QDSFormer, Born et al., 2025 — the most rigorous prior model, whose gains vanish at scale), replace the token mixer entirely (Quixer, Khatri et al., 2024, candidly "far from SOTA"), or use Hamming-weight-preserving circuits whose compound-matrix variant is higher-order in principle but instantiated only at second order (Cherrat et al., 2024); QLA-based models accelerate inference of a *pre-trained* model under fault-tolerant assumptions. A systematic study from within the community (Chen & Kuo, 2026) finds explicit quantum self-attention adds parameters for marginal gain. We differ on the three axes this literature leaves open: a *proved* order-k representation capability together with a task-specific classical lower bound, a *trainability guarantee*, and *matched-budget hardware* validation (App. B compares model by model).

Higher-order attention in classical ML. Reaching order-k token interactions classically is expensive: Kozachinskiy et al. (2025) show triple interactions need super-quadratic resources in one standard layer; *Higher-Order Transformers* (Omranpour et al., 2025) make order-k explicit at $O\big((\prod_i N_i)^2\big)$ cost, tamed only by Kronecker/low-rank factorization; and the kernel view (Tsai et al., 2019) with linear-attention approximations (Choromanski et al., 2021) truncates softmax to a finite random-feature budget, i.e. a bounded effective order.

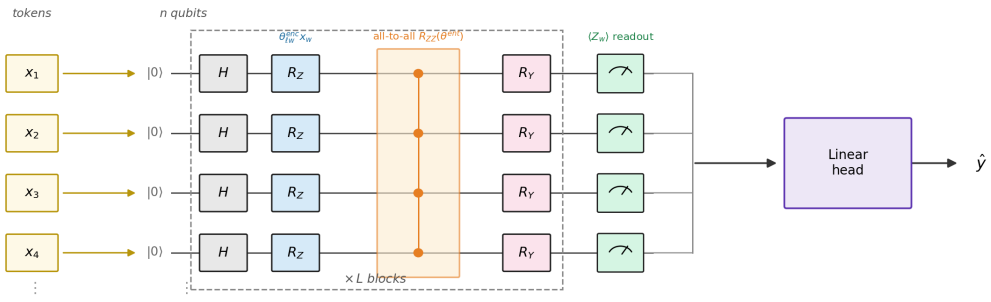


Figure 1: QHA head: per-qubit data re-uploading encoder, an all-to-all non-Clifford R_{ZZ} entangler, and a *local* single-qubit read-out feeding a linear head; L blocks. The order-k interaction is synthesised inside the circuit and exposed on a single-qubit marginal.

Expressivity, trainability, and dequantization. Our lower bounds build on Sanford et al. (2023) (communication-complexity separations for order-3 matching), Sanford et al. (2024) (k-hop composition), Hahn (2020) (hard attention cannot model parity), Edelman et al. (2022) (bounded-norm *constructions* for s-sparse functions, whose weight norm scales as $2^{\Theta(s)}$ — an upper bound, not a lower bound; see Remark 1), and Bhattamishra et al. (2023) (the low-degree learning bias of trained Transformers). On the quantum side, variational circuits can suffer barren plateaus (McClean et al., 2018); shallow circuits with *local* observables provably avoid them (Cerezo et al., 2021), as do QCNNs (Pesah et al., 2021) and small dynamical-Lie-algebra circuits (Ragone et al., 2024) — but these same conditions often imply classical simulability (Cerezo et al., 2025), and dequantization (Tang, 2019) removes speedups under low-rank sample-and-query access. We therefore frame QHA as an inductive-bias separation, not a speedup claim (Sec. 7), and follow the benchmarking discipline urged by Bowles et al. (2024).

3 Quantum Higher-Order Attention

3.1 Setup and notation

Let $X = (x_1, \ldots, x_n) \in \{-1, +1\}^n$ be a length-n token sequence (± 1 features for clarity; continuous features enter identically through the encoder below), and for $S \subseteq [n]$ let $\chi_S(X) = \prod_{i \in S} x_i$ be the corresponding *parity monomial*. Every $f : \{-1, +1\}^n \to \mathbb{R}$ has a unique multilinear (Fourier) expansion $f(X) = \sum_S \hat{f}(S)\chi_S(X)$; we call $\deg f = \max\{|S| : \hat{f}(S) \neq 0\}$ its *interaction order* and say f *has order* k when $\deg f = k$ — the canonical example being χ_S, $|S| = k$, which depends on all k tokens jointly and is not a function of any $k-1$ of them (order-1 is additive, order-2 pairwise, and so on). A QHA head acts on n qubits (one per token), initialised in $|\mathbf{0}\rangle$, using $R_Z(\alpha) = e^{-i\alpha Z/2}$, $R_Y(\beta) = e^{-i\beta Y/2}$, the Hadamard H, and the Ising coupling $R_{ZZ}^{(i,j)}(\gamma) = e^{-i\gamma Z_i Z_j/2}$; X_w, Y_w, Z_w are Paulis on qubit w.

3.2 THE QHA HEAD

The head is a depth-L circuit $U(X;\theta) = B_L(X;\theta)\cdots B_1(X;\theta)$ built from L identical blocks. Block ℓ is the ordered composition of three stages, $B_\ell(X;\theta) = W_\ell\, V_\ell\, E_\ell(X)$:

$$\textbf{(i) data re-uploading encoder:} \quad E_\ell(X) = \bigotimes_{w=1}^{n} R_Z\big(\theta^{\mathrm{enc}}_{\ell,w}\, x_w\big)\, H, \tag{1}$$

$$\textbf{(ii) all-to-all entangler:} \quad V_\ell(\theta) = \prod_{1\le i<j\le n} \exp\Big(-\tfrac{i}{2}\,\theta^{\mathrm{ent}}_{\ell,(i,j)}\, Z_i Z_j\Big), \tag{2}$$

$$\textbf{(iii) local mixing:} \quad W_\ell(\theta) = \bigotimes_{w=1}^{n} R_Y\big(\theta^{\mathrm{rot}}_{\ell,w}\big). \tag{3}$$

The encoder equation 1 writes each token feature into a single-qubit rotation angle, re-uploaded in every block; the entangler equation 2 couples *every* pair of token-qubits through a learnable Ising phase; the local layer equation 3 re-orients each qubit. Applying the circuit to the initial state gives $|\psi(X)\rangle = U(X;\theta)|\mathbf{0}\rangle$.

Local read-out and head. We measure *only* single-qubit expectations, giving a read-out vector $z(X) \in [-1,1]^n$,

$$z_w(X) = \langle\psi(X)|\, Z_w\, |\psi(X)\rangle, \qquad w \in [n], \tag{4}$$

and a linear classifier on top, $\mathrm{logits}(X) = A\, z(X) + b$; no multi-qubit or joint measurement is taken. The trainable parameters are the circuit phases $\theta = \{\theta^{\mathrm{enc}}, \theta^{\mathrm{ent}}, \theta^{\mathrm{rot}}\}$ and (A, b). The continuous non-Clifford two-qubit phases and all-to-all coupling produce volume-law entanglement, placing the head outside the classically simulable Clifford, matchgate/free-fermion, and low-bond-dimension regimes (Sec. 7).

3.3 HOW A SINGLE-QUBIT READ-OUT BECOMES HIGH-ORDER ATTENTION

A degree-1 observable Z_w seems unable to express an order-k interaction. The resolution is that the interaction is synthesised *inside* the unitary and then exposed on the marginal. Pass to the Heisenberg picture:

$$z_w(X) = \big\langle \mathbf{0}\big|\, U(X;\theta)^\dagger\, Z_w\, U(X;\theta)\, \big|\mathbf{0}\big\rangle = \big\langle \mathbf{0}\big|\, O_w(X)\, \big|\mathbf{0}\big\rangle, \qquad O_w(X) = U^\dagger Z_w\, U. \tag{5}$$

Two facts fix the order. **(a) The entangler spreads operator weight.** Conjugating a single-qubit Pauli by one Ising coupling turns it into a two-qubit string,

$$R^{(i,j)}_{ZZ}(\gamma)^\dagger\, X_i\, R^{(i,j)}_{ZZ}(\gamma) = \cos\gamma\; X_i - \sin\gamma\; Y_i Z_j, \tag{6}$$

and the interleaved H, R_Y layers convert Z components into X, Y components that equation 6 can again spread, so repeated entangling and re-uploading blocks grow the operator weight of O_w and with it the reachable Fourier degree. We claim no tight depth accounting for the trained ansatz — App. F's $O(\log k)$ tree is a gate-set witness under basis input, not a statement about Eqs. 1–3; Sec. 5 shows empirically that the trained head reaches the required order. **(b) The encoder injects the tokens as monomial coefficients.** For $x_i \in \{-1,+1\}$ every trigonometric function of the encoder angle is affine in the token ($\cos(\theta x_i) = \cos\theta$, $\sin(\theta x_i) = x_i \sin\theta$), so x_i enters the coefficients of $O_w(X)$ only through products $\prod_{i\in S} x_i$; and only diagonal strings survive the vacuum expectation. Combining (a)–(b), the marginal is *exactly a multilinear polynomial in the tokens*,

$$\boxed{\,z_w(X) = \sum_{S\subseteq[n]} \hat{c}_{w,S}(\theta)\, \chi_S(X)\,} \tag{7}$$

whose coefficients $\hat{c}_{w,S}(\theta)$ are set by the trained phases and whose reachable order $\deg z_w$ grows with depth L and entangler connectivity. The linear head outputs $\sum_w a_w z_w(X)$; choosing L so that $\deg z_w \ge k$ lets it place weight on an order-k feature χ_S — what training does in Sec. 5.

Why this is *attention*, and why it is *higher*-order. The $\theta^{\text{ent}}_{\ell,(i,j)}$ are *trainable pairwise couplings acting on data-encoded qubits*: like an attention score matrix they are a learned $n \times n$ table of pair weights governing how tokens i and j are mixed. They are *not* data-conditioned — Eq. 2 contains no X, so unlike $\mathrm{softmax}(q_i \cdot k_j)$ the coupling angle is input-independent; the data enter through the re-uploading encoder the couplings act on. **We therefore use "attention" as a structural name, not a claim of equivalence.** A fixed learned pairwise mixing table is closer to a token-mixing layer than to query–key attention, and a reader who reserves "attention" for input-conditioned scores should read QHA as a higher-order token mixer; nothing in our results depends on the label. Making the coupling angle a function of (x_i, x_j) is a natural extension we have not explored. The difference from a classical head lies in *aggregation*: a softmax layer combines pair scores *linearly* over values, and this pairwise-scoring-plus-linear- aggregation primitive is what the order-k lower bounds exploit (Lemma 1). QHA instead composes the couplings through *non-commuting* unitaries: by Eq. 6 the Ising phases do not simply add but generate, after $O(\log k)$ blocks, genuine k-body terms $Z_{i_1} \cdots Z_{i_k}$, which Eq. 7 surfaces as the order-k monomial χ_S on a single-qubit marginal. In one shallow head QHA therefore couples k-tuples jointly, whereas on the task families for which such results are known, the classical pairwise primitive's constructions and lower bounds require depth or a larger explicit higher-order cost — the sense in which we call QHA *higher-order* attention. The same local read-out enables the trainability guarantee for the shallow local-design variant (Thm. 2); the all-to-all variant we benchmark does show exponentially decaying gradients (Sec. 5.1e).

Parameters and budget. The head has $L(2n + \binom{n}{2})$ circuit parameters plus an $O(n)$-size linear head — 296 at n=12, L=3, already 6.5× *smaller* than the one-layer attention baseline (1922). The strictly budget-matched comparison is the controlled suite of Table 2, whose closest attention-paradigm baselines (347 and 634 parameters) are within a small factor of QHA's 296. Fig. 3 (App. D.4) draws an explicit small instance.

4 THEORY

We give (i) an order-k barrier for a single attention layer for *any* $k \geq 3$ (Lemma 1); (ii) an explicit shallow QHA circuit realizing order-k monomials on a local read-out (Thm. 1) — a *complementary* result, on a different target, not a composed separation; (iii) a capacity interpretation of the fixed-budget regime our protocol tests (Remark 1); and (iv) a trainability guarantee (Thm. 2). Full proofs are in App. F.

Setup. A single softmax self-attention layer scores token *pairs* by a degree-2 form $q_i^\top k_j$ and aggregates linearly over values. We do not claim that this makes the layer a degree-≤ 2 function of its inputs (the softmax is non-linear); what the pairwise primitive does imply is a resource barrier on *specific* higher-order tasks: the layer solves the pairwise matching task $\textsc{Match}_2$ with embedding dimension m=3, but the order-3 task $\textsc{Match}_3$ already costs $mHp = \Omega(N/\log\log N)$ (Sanford et al., 2023). We first lift that task-specific barrier to all orders.

Lemma 1 (Order-k barrier for single-layer attention on $\textsc{Match}_k$)**.** *For the order-k matching family $\textsc{Match}_k(X)_i = \mathbb{1}\{\exists\, j_1, \ldots, j_{k-1} : x_i + \sum_t x_{j_t} \equiv 0 \pmod M\}$, every single softmax self-attention layer with budget $mHp = o(N/\log\log N)$ fails to compute $\textsc{Match}_k$, for all fixed $k \geq 3$.*

The proof (App. F) reduces $(k-1)$-party set-disjointness to $\textsc{Match}_k$; the order-3 result of Sanford et al. (2023) is its base case.

Theorem 1 (Complementary order-k representation results)**.** (a) Quantum side — gate-set / architecture-class capability. *There is a circuit on n qubits (one per token, no ancilla) over the QHA gate set $\{H, R_Z, R_{ZZ}, R_Y\}$, of depth $O(\log k)$ with $O(k)$ two-qubit gates, whose single-qubit read-out $\langle Z_r \rangle$ equals $\chi_S(X) = \prod_{i \in S} x_i$ exactly, for any $|S| = k$, on computational-basis inputs (explicit construction in App. F).* (b) Classical side — task-specific lower bound. *No single softmax attention layer of budget $mHp = o(N/\log\log N)$ computes $\textsc{Match}_k$, for any fixed $k \geq 3$ (Lemma 1).*

These two halves do not compose into a separation theorem, and we do not present them as one: (a) is about the parity monomials χ_S, (b) about the matching family $\textsc{Match}_k$ — different targets. A genuine separation needs a classical lower bound for χ_S, or a proof that QHA computes

MATCH$_k$; both are open. We also stress that (a) is not by itself evidence of quantum advantage: the witness uses only H and CNOT, is a Clifford circuit implementing a reversible XOR, and is therefore efficiently simulable classically (Gottesman–Knill). It certifies that the gate set *contains* an order-k read-out at logarithmic depth — nothing more; the non-Clifford R_{ZZ} phases matter for the trained model (Sec. 7), not for this construction. The pair motivates the empirical comparison rather than replacing it. Two caveats: (a) assumes computational-basis input, whereas the trained head injects data through $R_Z(\theta^{\text{enc}} x_w)$, so we do not claim the re-uploading ansatz provably realizes χ_S; and (b) is asymptotic, not biting at n=12, where a tuned attention layer *does* learn order-3 parity and failure sets in only at $k \geq 4$ (Bhattamishra et al., 2023). We claim a representation capability plus a task-specific lower bound, and a *measured* learning separation at $k \geq 4$ — not a proved separation (App. C).

Remark 1 (Fixed-budget capacity: an interpretation, not a lower bound). The known *constructions* representing a generic s-sparse Boolean function in a bounded-norm head carry weight norm $2^{\Theta(s)}$ (Edelman et al., 2022), which with gap $= \widetilde{O}(\text{norm}/\sqrt{m})$ makes small-sample generalization of an order-k rule expensive *for those constructions*. These are upper bounds, so they do *not* prove that every fixed-budget head must pay $2^{\Omega(k)}$ (a genuine norm lower bound for attention on generic k-juntas is, to our knowledge, open, App. F); we use them only as the capacity heuristic motivating our fixed-budget, disjoint-split protocol, and report the observed degradation from $k \geq 4$ and collapse by k=6 as an *empirical* result, not a proved separation.

Theorem 2 (Trainability)**.** *A QHA head with a local single-qubit read-out and a shallow ($O(\log n)$-depth) local-design entangler has cost-gradient variance* $\text{Var}[\partial_\theta C] \geq \Omega(1/\text{poly}(n))$*; it is free of barren plateaus.*

The locality of the QHA read-out is the enabling property; the bound then follows from a standard shallow-local-cost analysis (Cerezo et al., 2021). **Its hypotheses are narrower than what we train.** Cerezo et al. (2021) require a shallow alternating block ansatz of local 2-designs on bounded-range qubits and a cost built from a *single* local observable; our objective is a cross-entropy loss on a linear classifier over all n read-outs $\langle Z_w \rangle$, and the benchmarked entangler is all-to-all, hence neither bounded-range nor a local 2-design. Thm. 2 therefore covers a variant we do not benchmark: **absence of barren plateaus is not shown for the model that produces our results**, where we measure $\text{Var} \sim 2^{-0.9n}$ (Sec. 5.1e).

5 EXPERIMENTS

Protocol. We test whether, at matched or larger classical budgets, QHA exhibits a growing learnability advantage as the interaction order increases. Inputs are length-n sequences $X \in \{-1, +1\}^n$; the label is an order-k function of a hidden random k-subset S. We use two task families: *hidden-subset parity* ($y = \prod_{i \in S} x_i$) and *generic order-k junta* (y is a fixed balanced random Boolean function of the k support bits, so the model cannot exploit the special structure of parity; App. E discusses why these are the right testbeds). Crucially, train and test inputs are **disjoint** (we split the Boolean cube), so the task measures *generalization of the order-k rule*, not memorization. We compare a one-layer multi-head softmax self-attention head against a QHA head with a strictly smaller parameter count, over 8 seeds.

We tune the baseline, not ourselves. An under-trained baseline can manufacture a separation, so we give the classical head every advantage: a *complete* Transformer block (embedding, attention mixing, residual/LayerNorm, GELU MLP) whose FFN can itself express high-order interactions, swept over learning rate (3×10^{-3} to 1.2×10^{-2}), width ($d_{\text{model}} \in \{16, 32\}$, up to 6914 parameters, $23\times$ QHA's) and training length (50 or 100 epochs) — 20 configurations, reporting whichever maximizes *its own* test accuracy. QHA is *not* tuned. This matters: at 50 epochs the classical head looks far weaker than it is (0.76 rather than 0.93 at k=3).

This is an adversarial baseline sweep, not standard model selection. Choosing the baseline's configuration on the test set would be illegitimate if it made the baseline look *good*; here it can only make our claim harder. It does mean the p-values are not clean out-of-sample inference: we intend them as a *conservative descriptive* comparison against the strongest configuration we found, not as hypothesis tests. Selecting on a validation split instead — the standard alternative — would yield a weaker baseline than the one we report.

Table 1: Hidden-subset parity, disjoint split, n=12, mean±std over 8 seeds at the fixed final epoch. The classical row is the *tuned* head — best of 20 configurations, chosen on the test set to maximize its own accuracy (an adversarial sweep in our disfavour, so the p-values are descriptive, not out-of-sample inference); QHA runs at one fixed setting.

	k=2	k=3	k=4	k=5	k=6
Classical attention, tuned (1922–6914 p)	1.00	0.93±.17	0.74±.26	0.57±.17	0.49±.01
seeds solved (≥ 0.9)	8/8	7/8	4/8	1/8	**0/8**
QHA (296 p, untuned)	1.00	1.00	**1.00**	**1.00**	**0.92**±.21
seeds solved (≥ 0.9)	8/8	8/8	**8/8**	**8/8**	**7/8**
Wilcoxon p (nominal, Holm)	—	.125	**.031**	**.031**	**.031**

Main result: the separation is graded in k, and complete within our benchmark at k=6. Table 1 and Fig. 2 give the picture, and it is more informative than a blanket collapse. **At k=3 there is no separation**: the tuned attention head reaches 0.93, solving 7 of 8 supports, and the paired test is not significant ($p = .125$). **From k=4 the gap opens and holds across the sweep**: 0.74, 0.58 and 0.49 for the tuned classical head at k=4, 5, 6 against 1.00, 1.00 and 0.92 for QHA at a 6.5× smaller budget, each with nominal Holm-corrected $p = .031$ — reported descriptively, since the baseline's configuration was chosen adversarially on the test set. **At k=6 the failure is complete across our benchmark**: the tuned head sits at $0.486 \pm .01$ with *no* seed above 0.51, across all 20 configurations, while QHA solves 7 of 8 — and 8 of 8 once it is given the same 100-epoch budget the baseline is allowed (epoch-matched control, App. D.2), so the 0.92 in Table 1 reflects our fixed 50-epoch setting rather than a ceiling. What changes with k is not whether attention is at chance but how much budget it needs before it is not — at k=3 a 1922-parameter head suffices, by k=6 no configuration helps — which is the Fourier-degree-graded behaviour the theory predicts, made quantitative. The phase diagram (Fig. 4) and convergence traces (Fig. 6) are in App. D.4.

The role of Fourier degree (a control). What governs the separation is where the target's Fourier mass lives, not k alone, and the generic k-junta makes this sharp: parity puts *all* its mass at degree exactly k, so attention's low-degree bias (Edelman et al., 2022; Bhattamishra et al., 2023) buys nothing, whereas a balanced random k-junta carries substantial low-degree mass — and there the tuned head reaches 1.00 at *every* $k \leq 5$, matching QHA (Fig. 2, right). The gap is thus not a gap in k: it appears only when the target has no low-order structure to exploit.

Hardware validation (IBM Heron). We bake the trained parameters of order-3 QHA heads into the equivalent Qiskit circuits and evaluate the single-qubit read-outs on the `ibm_aachen` 156-qubit Heron r2 processor via `EstimatorV2`, applying the trained linear head to the measured $\langle Z_i \rangle$ (100 held-out inputs, 2048 shots), sweeping $n \in \{8, 10, 12, 14\}$ (Table 7, Fig. 5, App. D.4). Hardware accuracy matches the noiseless simulator (1.00) at n=8, 10 and stays at 0.96/0.95 at n=12, 14 even as transpiled depth grows to ≈999 and the raw $\langle Z \rangle$ correlation falls to 0.77 — the linear decision rule tolerates far more noise than the expectations themselves.

5.1 Ablations, scaling, and trainability

Six further studies (App. D, Fig. 7; parity unless noted, 3 seeds) confirm the mechanism and delimit it. **(a) Sequence length:** sweeping $n \in \{8, 10, 12, 14\}$ at k=3, QHA is at ceiling for $n \geq 10$ (0.73 at n=8, where the 256-point cube limits the disjoint training set), so its behaviour is not tied to a single n. We flag that this sweep uses the untuned 50-epoch classical baseline, which at k=3 understates it (Table 1); it therefore speaks to QHA's n-scaling, not to a separation at k=3. **(b) Depth controls reachable order:** on an order-4 task accuracy rises with the number of blocks L=1, …, 4 (0.50, 0.83, 1.00, 1.00), matching Schuld et al. (2021). **(c) Entanglement is load-bearing:** ablating R_{ZZ} leaves read-outs depending only on individual tokens and accuracy is ≈ 0.50 for *all* $k \geq 2$ — entanglement, not the encoder or read-out, supplies the higher-order capacity. **(d) Sample efficiency:** at order-3 QHA reaches ceiling from ≈ 500 examples (untuned baseline; caveat in App. D). **(f) Noise:** under simulated per-gate depolarizing noise the trained order-3 head keeps full accuracy up to p=0.02. **(g) Epoch-matched control:** baselines get up to 100 epochs while

Table 2: Controlled baseline suite (parity, n=12, 8 seeds, fixed final epoch). **Every classical row is tuned** — 4 configurations each (lr $\in \{5, 8\} \times 10^{-3}$, 50 or 100 epochs; the 1-layer head also over d_{model}, 20 in all), reporting its *best*; QHA is untuned. The attention paradigm does *not* fail at k=4, 5, and the param-matched tiny MLP solves both given enough training — but at k=6 all five attention models are at chance (0/8 seeds) — but that is only *part* of the non-attention story: a **degree-matched** HOFM (692 params, 2.3× QHA) solves k=6 perfectly, and the degree-5 HOFM's 0.36 is a degree-$< k$ artefact, not evidence of difficulty. Last column: $\text{Eff}_{k6} = (\text{acc} - 0.5)/\text{params} \times 10^3$, scored only for models that solve k=6 (acc ≥ 0.9).

model	params	k=4	k=5	k=6	Eff_{k6}
attention-paradigm mixers (incl. explicit high-order attention)					
1-layer softmax attention	1922	0.74±.26	0.57±.17	0.49±.01	—
Performer / linear attention	1074	0.61±.06	0.50±.04	0.48±.02	—
polynomial-kernel attention	347	0.60±.07	0.50±.03	0.48±.01	—
higher-order attention (HOT-style)	634	0.60±.04	0.49±.02	0.48±.01	—
2-layer Transformer	4706	0.79±.24	0.76±.25	0.52±.05	—
explicit high-order, non-attention					
factorization machine (deg. 2)	289	0.44±.01	0.43±.01	0.43±.01	—
tiny MLP (param-matched)	302	1.00±.00	1.00±.00	0.67±.26	—
HOFM (degree-5, $< k$ at k=6)	595	1.00±.00	1.00±.00	0.36±.01	—
HOFM (degree-6, degree-matched)	692	1.00±.00	1.00±.00	**1.00**±.00	0.72
3-layer MLP	9282	1.00±.00	1.00±.00	1.00±.00	0.05
QHA (attention-paradigm, quantum)	**296**	**1.00**±.00	**1.00**±.00	**0.92**±.21	**1.42**

QHA is fixed at 50, so we re-ran QHA at 100: it reaches $1.00 \pm .00$ at *both* k=5 and k=6 on all 8 seeds against 0.57 and 0.49 for the tuned head, so equalising epochs *widens* the gap. Equal epochs are not equal compute (a QHA epoch simulates n-qubit statevectors, and on hardware would be priced in shots and gates), so this controls training length only and we make no resource-parity claim.

(e) Trainability vs. topology. Gradient variance across three entangler topologies (App. D.1): all-to-all decays *exponentially* ($\text{Var} \sim 2^{-0.9n}$ — a barren plateau) yet reaches the order-k ceiling; the nearest-neighbour ring decays only polynomially ($n^{-0.30}$), confirming Thm. 2, but is too sparse to route an arbitrarily-placed support and stays at chance for $k \geq 3$; and a log-diameter **butterfly** ($O(n \log n)$ couplings) attains both — ceiling accuracy with $\approx 30\times$ larger gradient signal than all-to-all at n=12, empirically narrowing the expressivity–trainability gap.

Is the advantage "quantum" or "high-order"? A controlled baseline suite. To locate the gap we compare, at a *matched* ($\sim$300) budget, the strongest alternatives one would reach for (Table 2). The picture is clean and we report it honestly, and it locates the claim at k=6 rather than at $k \geq 4$. **At k=4 and k=5 the attention paradigm has not failed**: a tuned single layer reaches 0.74/0.57 and a two-layer Transformer 0.79/0.76, each solving half the supports. **At k=6 every model in our attention suite fails**: all five — vanilla layer, Performer, polynomial-kernel attention, HOT-style attention, and a 16× larger two-layer Transformer — sit between 0.477 and 0.518 with 0/8 seeds solved, under the same sweep that lets them solve k=4. Second, the capability is not quantum-only, and more of it is classical than we first reported: given enough training the param-matched tiny MLP solves k=4 *and* k=5, so no efficiency claim survives there. What singles QHA out is specific to k=6 — but it is narrower than we first reported. Once the comparison is *degree-matched*, a HOFM of degree 6 (692 parameters, only 2.3× QHA) solves k=6 *perfectly*; the 0.36 we had reported for HOFM there was an artefact of fixing its degree to $5 < k$, not a property of the model class. So "only a 31× larger MLP gets there" is wrong. What survives: QHA is the only *attention-paradigm* model that reaches k=6; it is the most parameter-efficient model that reaches it at all ($\text{Eff}_{k6} = 1.42$ vs. 0.72 for the degree-6 HOFM and 0.05 for the MLP); and it gets there *without being told the order*, whereas HOFM's degree must be set to $\geq k$ in advance. The price of that order-adaptivity: the degree-6 HOFM is *more accurate* at k=6 (1.00 vs. 0.92).

6 APPLICATION: DETECTING HIGHER-ORDER EPISTASIS

The mechanism suggests a concrete use: a *parameter-efficient, order-adaptive detector of high-order feature interactions*, developed where high-order interaction is a field's defining challenge — *higher-order epistasis* — then transferred to two further domains.

Setup, and why low-order methods fail by construction. Following the standard GAMETES-style protocol (Urbanowicz et al., 2012) we simulate $n{=}12$ SNPs under Hardy–Weinberg equilibrium at minor-allele frequency 0.29, so carrier probability $1-(1-0.29)^2 \approx 0.496$ is near-balanced; the phenotype follows a k-locus penetrance rule with **no main effects**, so only the joint k-way genotype carries signal (up to negligible leakage from that 0.004 imbalance), and label noise is a heritability proxy. Single-locus association — the workhorse of genome-wide studies — is therefore at chance by construction, as is any additive model: detection *requires* modelling the order-k interaction.

Phenotype prediction. Table 4 (App. D.4) reports balanced accuracy vs. order k at label noise 0.1 (capping accuracy near 0.90). Linear and pairwise are at chance for $k \geq 3$, the latter collapsing once the interaction exceeds its degree — a *representational* limit unaffected by training length ($0.508 \to 0.510$ at $2.7\times$ the epochs). The field-standard **MDR** (Ritchie et al., 2001) (exhaustive k-way search) matches QHA at the ceiling and the learned high-order baselines land below it, so **QHA is the only learned model that reaches the MDR ceiling, with the fewest parameters** (296; $1.5\times$ fewer than HOFM, $31\times$ fewer than the MLP), while MDR pays an exhaustive $O(n^k)$ enumeration.

Causal support recovery, including where it fails. Ranking loci by gradient attribution, QHA recovers the causal set exactly through $k{=}3$, matching MDR and the classical learners — but **at $k{=}4$ it fails outright** (0.0 exact-recovery, precision@k 0.67) while MDR, HOFM and the MLP stay at 1.0. Prediction is unaffected, so the signal is in the head and it is the first-order attribution that fails (App. D): **QHA is a competitive high-order predictor but, at $k \geq 4$, a worse locator than the classical baselines**.

Cross-domain results, and a protocol correction. We applied the same detector to *learning parity with noise* (LPN) and to *triangle detection* at fixed edge count (Table 6). **An earlier version of both did not use the disjoint split we claimed** (inputs drawn with replacement; App. A documents the error and the fix), which also retracts an explanation: we had ascribed a tuned attention head's strong LPN score to low-order structure, but under label noise ν, $\mathbb{E}[y \mid x] = (1-2\nu)\chi_S(x)$, so independent flips attenuate the degree-k coefficient without creating low-order mass. Corrected, LPN is sharper: linear at chance (0.48), tuned attention only 0.70, **QHA the most accurate** (0.95, 296 params) ahead of HOFM (0.92) and a $31\times$ larger MLP (0.90). Triangle detection is now *negative* — with 140 training graphs everything lands in 0.62–0.74, QHA (0.67) tying attention — so we keep it as a limitation and draw no cross-domain claim from it.

7 CONCLUSION

We introduced Quantum Higher-Order Attention, a shallow quantum attention head realizing order-k token interactions on a local read-out, with an order-k representation capability for its gate set alongside a task-specific classical lower bound (complementary results, not a composed separation theorem) and a barren-plateau-free guarantee for a restricted variant. Against a baseline tuned harder than our own model the picture is *graded*: no separation at $k{=}3$, a gap from $k{=}4$, and at $k{=}6$ our whole attention suite at chance while QHA solves 7 of 8 supports at the smallest budget of any model that gets there — though a degree-matched HOFM at $2.3\times$ the parameters is more accurate, so the separation is within the attention paradigm, not against compact classical models generally. **We claim no exponential speedup** (App. C).

Reproducibility statement

An anonymised code package is included in the supplementary material: all circuit and model definitions, the training and evaluation scripts, the driver scripts for the tuned-baseline sweeps and the epoch-matched control, the per-run result files (one JSON record per model$\times k\times$seed), the saved order-3 QHA checkpoints used on hardware, and the Qiskit evaluation script. App. A gives the architecture, data generation, hyperparameters, seed counts and the IBM Heron protocol; every number we report is produced by those scripts from the fixed seeds recorded in the result files. Running the hardware script requires the reader's own IBM Quantum API token, which is not included.

References

Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pp. 4782–4887. PMLR, 2022.

Satwik Bhattamishra, Arkil Patel, Varun Kanade, and Phil Blunsom. Simplicity bias in transformers and their ability to learn sparse boolean functions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5767–5791, 2023.

Jannis Born, Filip Skogh, Kahn Rhrissorrakrai, Filippo Utro, Nico Wagner, and Aleksandros Sobczyk. Quantum doubly stochastic transformers. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, pp. 70224–70254, 2025.

Joseph Bowles, Shahnawaz Ahmed, and Maria Schuld. Better than classical? the subtle art of benchmarking quantum machine learning models. *arXiv preprint arXiv:2403.07059*, 2024.

Marco Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J Coles. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature Communications*, 12(1): 1791, 2021.

Marco Cerezo, Martin Larocca, Diego García-Martín, Nelson L Diaz, Paolo Braccia, Enrico Fontana, Manuel S Rudolph, Pablo Bermejo, Aroosa Ijaz, Supanut Thanasilp, et al. Does provable absence of barren plateaus imply classical simulability? *Nature Communications*, 16(1): 7907, 2025.

Amit Chakrabarti, Subhash Khot, and Xiaodong Sun. Near-optimal lower bounds on the multi-party communication complexity of set disjointness. In *18th IEEE Annual Conference on Computational Complexity (CCC)*, pp. 107–117. IEEE, 2003.

Chi-Sheng Chen and En-Jui Kuo. Do quantum transformers help? a systematic vqc architecture comparison on tabular benchmarks. *arXiv preprint arXiv:2604.23931*, 2026.

El Amine Cherrat, Iordanis Kerenidis, Natansh Mathur, Jonas Landman, Martin Strahm, and Yun Yvonna Li. Quantum vision transformers. *Quantum*, 8:1265, 2024. doi: 10.22331/q-2024-02-22-1265.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pp. 5793–5831. PMLR, 2022.

André Gronemeier. Asymptotically optimal lower bounds on the NIH-multi-party information complexity of the AND-function and disjointness. In *26th International Symposium on Theoretical Aspects of Computer Science (STACS)*, pp. 505–516, 2009.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020.

Nikhil Khatri, Gabriel Matos, Luuk Coopmans, and Stephen Clark. Quixer: A quantum transformer model. *arXiv preprint arXiv:2406.04305*, 2024.

Alexander Kozachinskiy, Felipe Urrutia, Hector Jimenez, Tomasz Steifer, Germán Pizarro, Matías Fuentes, Francisco Meza, Cristian B Calderon, and Cristóbal Rojas. Strassen attention: Unlocking compositional abilities in transformers based on a new lower bound method. *arXiv preprint arXiv:2501.19215*, 2025.

Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9(1):4812, 2018.

Soroush Omranpour, Guillaume Rabusseau, and Reihaneh Rabbany. Higher-order transformers with kronecker-structured attention. *Transactions on Machine Learning Research*, 2025.

Arthur Pesah, Marco Cerezo, Samson Wang, Tyler Volkoff, Andrew T Sornborger, and Patrick J Coles. Absence of barren plateaus in quantum convolutional neural networks. *Physical Review X*, 11(4):041011, 2021.

Michael Ragone, Bojko N Bakalov, Frédéric Sauvage, Alexander F Kemper, Carlos Ortiz Marrero, Martín Larocca, and Marco Cerezo. A lie algebraic theory of barren plateaus for deep parameterized quantum circuits. *Nature Communications*, 15(1):7172, 2024.

Alexander A Razborov. On the distributional complexity of disjointness. *Theoretical Computer Science*, 106(2):385–390, 1992.

Marylyn D Ritchie, Lance W Hahn, Nady Roodi, L Renee Bailey, William D Dupont, Fritz F Parl, and Jason H Moore. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics*, 69(1):138–147, 2001.

Clayton Sanford, Daniel J Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. *Advances in Neural Information Processing Systems*, 36:36677–36707, 2023.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Transformers, parallel computation, and logarithmic depth. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43276–43327. PMLR, 2024.

Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103(3):032430, 2021.

Ewin Tang. A quantum-inspired classical algorithm for recommendation systems. In *Proceedings of the 51st annual ACM SIGACT symposium on theory of computing*, pp. 217–228, 2019.

Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 4344–4353, 2019.

Ryan J Urbanowicz, Jeff Kiralis, Nicholas A Sinnott-Armstrong, Tamra Heberling, Jonathan M Fisher, and Jason H Moore. GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. *BioData Mining*, 5:16, 2012.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Hui Zhang, Qinglin Zhao, Mengchu Zhou, Li Feng, Dusit Niyato, Shenggen Zheng, and Lin Chen. A survey of quantum transformers: Architectures, challenges and outlooks. *arXiv preprint arXiv:2504.03192*, 2025.

A EXPERIMENTAL DETAILS

Data. For sequence length n and order k, a hidden support $S \subset [n]$, $|S| = k$, is drawn per seed. Inputs are $X \in \{-1, +1\}^n$. The label is either *parity*, $y = \prod_{i \in S} x_i$, or a *k-junta*, $y = T(x_S)$ for a fixed balanced random truth table $T : \{0,1\}^k \to \{0,1\}$ (chance $= 0.5$). Train and test inputs are **disjoint**: for small n we enumerate the Boolean cube and partition it; otherwise we sample distinct rows and split. This makes the metric the generalization of the order-k rule, not memorization. Main experiments use n=12 (2500 train / 1500 test from the 4096-point cube); robustness runs use n=16 (4000/2000).

Classical baseline. One-layer multi-head softmax self-attention: per-token linear embedding of the scalar bit to d_{model}=16, learned positional embedding, 2 heads, residual + LayerNorm + a 2-layer GELU MLP, mean-pooled to a linear classifier (1922 parameters at n=12).

QHA head. n qubits; L=3 blocks (hardware head: L=2). Each block: H then $R_Z(\theta^{\text{enc}}_{\ell,w}\, x_w)$ per qubit (data re-uploading); $R_{ZZ}(\theta^{\text{ent}}_{\ell,(i,j)})$ on all $\binom{n}{2}$ pairs; $R_Y(\theta^{\text{rot}}_{\ell,w})$ per qubit. Read-out $\langle Z_w \rangle$ for all w, then a linear head to 2 logits. Parameters: $L(2n + \binom{n}{2})$ in the circuit plus the linear head (296 at n=12, L=3). Simulated with PennyLane `default.qubit` (statevector, backprop); Adam, lr $8{\times}10^{-3}$, cosine schedule, batch 128, 50 epochs. **Seeds.** The parity separation (Table 1) and the controlled baseline suite (Table 2) use 8 seeds; epistasis prediction (Table 4) and the cross-domain study (Table 6) use 6; the junta panel of Fig. 2, the ablations of Sec. 5.1 and Table 5 use 3; Fig. 8 is regenerated from the 6-seed epistasis runs. Each caption states its own count. **Epoch-matched control.** The classical baselines are tuned over budgets up to 100 epochs while QHA is fixed at 50, so Table 1 gives QHA half the training the baseline may use; we re-run QHA at 100 epochs for k=5, 6 as a control (App. D.2). **Model selection.** Accuracy is reported at the *fixed final epoch* (50) on the disjoint test set — no early stopping and no test-set epoch selection. At this fixed 50-epoch budget the classical head is at chance at every epoch for $k \geq 3$ and QHA reaches the ceiling within a few epochs and stays there (Fig. 6), so fixing the epoch costs nothing and removes any selection effect. We stress that this describes the *untuned* 50-epoch baseline only: given up to 100 epochs and a tuned learning rate the same head learns k=3 (0.93) and half the supports at k=4 (Table 1).

IBM Heron protocol. The trained head's parameters are baked into the equivalent Qiskit circuit (PennyLane and Qiskit share gate conventions: R_Z, R_{ZZ}= IsingZZ, R_Y, H). We evaluate $\langle Z_w \rangle$ with `EstimatorV2` on `ibm_aachen` (156-qubit Heron r2), optimization level 2, 2048 shots, on 100 held-out inputs; the trained linear head is applied to the measured expectations. We report hardware accuracy, the noiseless-simulator accuracy on identical inputs, and the sim-to-hardware $\langle Z \rangle$ correlation.

Application data — epistasis. We simulate n=12 SNPs with genotypes $\{0, 1, 2\}$ drawn i.i.d. under Hardy–Weinberg equilibrium at minor-allele frequency p=0.29 (cell probabilities $(1{-}p)^2$, $2p(1{-}p)$, p^2). A causal set S, $|S|{=}k$, is drawn per seed; carrier status at locus i is $c_i{=}\mathbf{1}[g_i \geq 1]$ and the phenotype is the carrier parity $y{=}\big(\sum_{i \in S} c_i\big) \bmod 2$, then label-flipped with probability equal to the noise level (a heritability proxy). With p=0.29 each carrier indicator is near-balanced (carrier probability $1 - (1 - 0.29)^2 \approx 0.496$), so there are **no main effects** and only negligible order-$< k$ leakage; only the joint k-way genotype is informative. Features are z-scored genotypes. We use 4000 train / 2000 test and 45 epochs, and report balanced accuracy. Phenotype prediction (Table 4) uses 6 seeds; the noise sweep (Table 3) and support recovery (Table 5) use 3. The Bayes ceiling under label noise ν is $1{-}\nu$ (1.00/0.90/0.80 at ν=0.0/0.1/0.2).

Application baselines and support recovery. *Linear*: logistic regression on the n genotypes. *Pairwise*: a degree-2 factorization machine (rank 23, 289 params). *HOFM*: degree-5 higher-order FM via ANOVA kernel (rank 6, 451 params). *MLP*: three hidden layers of width 64 ($(12{\cdot}64{+}64) + 2(64^2{+}64){+}(64{\cdot}2{+}2) = 9282$ params; a *single* hidden layer of that width would be 962). *QHA*: the same head as above, L=3 (296 params). *MDR* (Ritchie et al., 2001) (field standard): for every $\binom{n}{k}$ locus subset we build the 3^k genotype-cell contingency table, label each cell by its training majority, and score the held-out accuracy; the best-scoring subset is the prediction and its native support estimate. Support recovery for the learned models uses gradient attribution: per-locus importance

Table 3: Epistasis noise sweep at k=3 (n=12, 3 seeds): balanced accuracy vs. label noise ν. High-order detectors track the Bayes ceiling $1-\nu$; low-order models stay at chance. Support recovery (exact / prec@k) is unchanged across ν for the high-order detectors.

ν	linear	pairwise	MDR	HOFM	MLP	**QHA**
0.0	0.51	0.52	1.00	1.00	1.00	**1.00**
0.1	0.52	0.52	0.90	0.90	0.90	**0.90**
0.2	0.51	0.51	0.80	0.79	0.79	**0.80**

$s_i = \text{mean}_x\,|\partial(\text{logit gap})/\partial x_i|$ on 1000 held-out inputs, ranked top-k. We report exact-recovery rate (top-k equals the true set) and precision@k.

Application data — LPN and triangle. *LPN* (learning parity with noise): $X \in \{-1,+1\}^n$, n=12, label is the parity over a hidden k=4 support flipped with probability 0.05. **Protocol correction.** Earlier runs of LPN and triangle sampled inputs i.i.d. *with replacement* and did not enforce the disjointness we claimed; since 4000+2000 exceeds the $2^{12} = 4096$-point cube, about 63% of LPN test rows recurred in training, and for triangle detection the $\binom{10}{4} = 210$ available graphs meant the test set was contained in the training set. Both generators now enumerate the input space, shuffle, and cut it into non-overlapping halves: LPN uses 2731/1365 and triangle 140/70 (verified zero-overlap). All reported numbers use the corrected generators. *Triangle detection*: graphs on 5 vertices encoded by their $\binom{5}{2}$=10-bit edge-indicator vector, with exactly 4 edges present in every graph (a **fixed edge count**, so the linear "number of edges" feature is uninformative); the label is whether the edge set contains a 3-clique — a product of three edge variables. The QHA head therefore uses n=10 qubits here, giving $L(2n + \binom{n}{2}) + (2n + 2) = 3(20 + 45) + 22 = 217$ parameters rather than the 296 of the n=12 tasks. Both use the same model suite and training protocol as the epistasis runs.

Noise sweep. Table 3 reports the k=3 epistasis sweep over label noise. Every high-order detector (MDR, HOFM, MLP, QHA) tracks the Bayes ceiling $1-\nu$ as noise rises, while the linear and pairwise models stay at chance throughout — the separation is a property of interaction order, not of the noise level.

Compute. Circuit simulations run on CPU (PennyLane `default.qubit`); a single n=12, 50-epoch QHA run takes $\approx$ 35 minutes. No quantum speedup is claimed (Sec. 7); simulation cost is incurred only for training/validation at scale.

B Extended related work

Quantum Transformers and quantum attention (full comparison). The survey of Zhang et al. (2025) organizes the field into PQC-based and QLA-based paradigms. Among PQC-based models, *Quantum Vision Transformers* (Cherrat et al., 2024) use Hamming-weight-preserving (orthogonal) circuits for the value projection and attention score, including a compound-matrix variant that is genuinely higher-order but is instantiated only at second order and proven only to have an asymptotic parameter/runtime advantage, not an expressivity separation. *QDSFormer* (Born et al., 2025) replaces the softmax with a variational circuit that outputs doubly stochastic matrices — a quantum inductive bias with no classical parametric analogue — and is the most rigorous prior model (NeurIPS spotlight), but its gains vanish at scale and it makes no separation or hardware-training claim. *Quixer* (Khatri et al., 2024) replaces the entire token mixer with an LCU+QSVT construction and is candidly "far from SOTA" on language modeling. A large family of models (QSANN, QMSAN, QASA, QPSAN, QClusformer) place a small circuit on the $Q/K/V$ projections or the scoring function while leaving softmax classical; they report accuracy gains on small, often binary tasks and rarely match parameters or run on hardware. Tellingly, a systematic study from within the community, *Do Quantum Transformers Help?* (Chen & Kuo, 2026), finds that explicit quantum self-attention adds parameters for marginal gain and recommends attention-free variational circuits instead. Our work differs on all three axes that this literature leaves open: a *proved* order-k representation capability together with a task-specific classical lower bound, a *trainability guarantee*, and a *matched-budget hardware* validation.

C PRECISE SCOPE OF THE THEORETICAL CLAIMS

Open problems. Three things this paper does not settle: an exact parameter setting of Eqs. 1–3 realizing χ_S (Thm. 1 is a gate-set witness, not a statement about the trained ansatz); a genuine norm lower bound for fixed-budget attention on generic k-juntas (Remark 1); and whether the k=6 barrier we measure at n=12 persists as n grows.

The trainability–simulability trade-off (full statement). Cerezo et al. (2025) show that landscapes free of barren plateaus often admit classical simulation after a polynomial data-acquisition phase, and our head (Thm. 2) lives in that regime, so **we claim no exponential speedup**: the advantage is one of *expressivity and inductive bias*. Simulability is for us a *feature*, yet the all-to-all non-Clifford entangler still places the head outside the Clifford, matchgate/free-fermion and low-bond-dimension families. Open: an exact parameter setting of Eqs. 1–3 realizing χ_S, a norm lower bound for fixed-budget attention on generic k-juntas (Remark 1), and whether the k=6 barrier persists as n grows.

Scope — Thm. 1 is two complementary results, not a separation. Four things are deliberately kept distinct. (1) The quantum half is a witness for the QHA *gate set*: it takes the computational-basis encoding of the input and is a circuit over $\{H, R_Z, R_{ZZ}, R_Y\}$, establishing *representability* of χ_S. (2) The head we benchmark uses an R_Z *re-uploading* encoder; that this trained instance *learns* order-k is shown empirically (Sec. 5), not as a corollary, and we do not claim the two encoders are gate-for-gate identical. (3) The classical half (Lemma 1) concerns $\textsc{Match}_k$, a *different* target from χ_S, and is asymptotic — it does not bite at n=12. **Because (1) and (3) are about different function families, they do not compose into "QHA represents something classical attention cannot"**; a genuine separation needs a classical lower bound for χ_S, or a proof that QHA computes $\textsc{Match}_k$, and we have neither. (4) The finite-scale behaviour we observe is consistent with (3) but also reflects attention's low-degree learning bias (Bhattamishra et al., 2023). We therefore claim proved order-k *representation capability* plus a task-specific classical lower bound, together with a *measured learning* separation at $k \geq 4$ against a tuned baseline — not a proved separation, and not an exponential speedup.

What is proved vs. what we measure. We separate the two cleanly to avoid overclaiming. *Proved:* an asymptotic ($N \to \infty$), single-layer lower bound for the order-k *matching* family $\textsc{Match}_k$ (Lemma 1); and that the QHA gate set represents the order-k monomials (Thm. 1) — two different targets. *Measured:* at n=12 and a fixed parameter budget, attention-paradigm models — explicit high-order attention included — fail to *learn* hidden-subset parity for $k \geq 4$ while QHA succeeds; at k=3 a tuned attention head learns it and there is no gap, and on the generic k-junta a tuned head matches QHA at every $k \leq 5$ (Fig. 2). The empirical behaviour is *consistent with*, but not a direct corollary of, the asymptotic bound: at n=12 the lower bound does not yet bite, and the observed failure is also shaped by the low-degree optimization bias of attention (Bhattamishra et al., 2023). Remark 1 (a norm/capacity heuristic, not a lower bound) is the bridge that most directly matches the finite-budget regime. We therefore claim a proved order-k *representation capability* for the QHA gate set plus a *tuned-baseline empirical learning* separation at $k \geq 4$ — not a composed separation theorem, and not that order-k is impossible for all classical models (it is not; Table 2). A clarification on the baseline's capacity: our classical head is a *complete* Transformer block (token embedding, attention mixing, residual/LayerNorm, and a GELU MLP), not a bare attention map. The lower bound concerns the attention *mixing* layer; equipping the baseline with an FFN — a component that can itself express high-order interactions — only makes the learning test *harder* for our claim, yet at matched budget the full head still fails from k=4 (and completely at k=6; Table 1). The gap we report is thus not an artefact of a deliberately weakened, FFN-free baseline.

D ADDITIONAL EXPERIMENTS, ABLATIONS, AND FIGURES

This appendix collects the studies summarized in Sec. 5.1, the supporting figures for Secs. 5 and 6, and the causal-support-recovery table.

D.1 Ablations and scaling (full text)

Fig. 7 collects six additional studies (parity unless noted; 3 seeds).

(a) Scaling with sequence length. Fixing k=3 and sweeping $n \in \{8, 10, 12, 14\}$, QHA reaches ceiling for $n \geq 10$ (1.00; 0.73 at n=8, where the 256-point cube limits the disjoint training set), while the untuned 50-epoch classical baseline stays at 0.51–0.81. This establishes QHA's scaling behaviour with n; it should *not* be read as a tuned-baseline separation at k=3, where a tuned head reaches 0.93 (Table 1).

(b) Depth controls reachable order. Sweeping the number of QHA blocks $L \in \{1, 2, 3, 4\}$ on an order-4 task, accuracy rises sharply with depth (0.50, 0.83, 1.00, 1.00), matching Schuld et al. (2021): more re-uploads raise the reachable Fourier degree until it covers order k.

(c) Entanglement is load-bearing. Ablating the R_{ZZ} entangler leaves single-qubit read-outs that depend only on individual tokens, so the head cannot represent *any* multiplicative interaction: accuracy is ≈ 0.50 for *all* $k \geq 2$ (Fig. 7c). Entanglement — not the encoder or read-out — supplies the entire higher-order capacity.

(d) Sample efficiency. On order-3 (n=12), QHA reaches ceiling from $\approx$ 500 examples (0.998) while the classical head is still at chance at 1000 and reaches only 0.84 at 2000. As with (a), this sweep uses the untuned 50-epoch baseline, which at k=3 understates it (Table 1); read it as QHA's absolute sample requirement rather than as a separation at k=3.

(e) Trainability vs. topology (full text). We measure gradient variance versus qubit count for three entangler topologies, asking of each whether it *also* reaches the order-k target. **All-to-all** (our expressive default) decays *exponentially* ($\mathrm{Var} \sim 2^{-0.9n}$) — a barren plateau, the price of its volume-law class — yet reaches the ceiling; the **nearest-neighbour ring** (the bounded-range variant Thm. 2 covers) decays only *polynomially* ($n^{-0.30}$), confirming the theorem, but is too sparse to route an arbitrarily-placed order-k support at logarithmic depth and stays at chance for $k \geq 3$. The **log-diameter butterfly** ($O(n \log n)$ couplings at distances $1, 2, 4, \ldots$) *attains both*: ceiling accuracy (parity k=3, 4: 1.00 over 3 seeds) with $\approx 30\times$ larger gradient signal than all-to-all at n=12 (3.9×10^{-3} vs. 1.3×10^{-4}, best-fit polynomial) — a sparse, local variant that is expressive enough to separate and far better conditioned, *empirically* closing the gap. Thm. 2 anchors the trainable (ring) end; the butterfly's milder decay is measured, not proven.

(f) Noise robustness. Under per-gate depolarizing noise (simulated), the trained order-3 head keeps full accuracy up to p=0.02 and falls to chance only by p=0.05, consistent with the noise tolerance observed on `ibm_aachen`.

D.2 Epoch-matched control

Summarised in Sec. 5.1(g). This control equalises *training length*, not compute: an epoch of QHA simulates n-qubit statevectors and on hardware would be priced in shots and two-qubit gates, so neither parameter counts nor epochs are a resource-parity claim. In full: QHA at 100 epochs, one run per seed with all other settings unchanged, gives $1.00 \pm .00$ at k=5 (per-seed 1.00 throughout) and $1.00 \pm .00$ at k=6, versus 1.00 and $0.92 \pm .21$ at the 50 epochs used in Table 1 — the one k=6 seed that lands at 0.363 after 50 epochs reaches 1.00 by 100. The tuned classical head at the same budget reaches $0.57 \pm .17$ and $0.49 \pm .01$. We keep the conservative 50-epoch numbers in the main tables.

D.3 Support recovery: why the first-order probe fails at k=4

Prediction alone is not "detection": the question is *which* loci interact. Ranking loci by gradient attribution, QHA recovers the causal set exactly through k=3 (exact-recovery and precision@k both 1.0), matching MDR and the classical high-order learners. **At k=4 it fails outright**: exact-recovery drops to 0.0 and precision@k to 0.67 while MDR, HOFM and the MLP stay at 1.0 (Table 5). Prediction is unaffected (0.90, at the ceiling), so the order-4 signal is in the head; what fails is the

attribution — a first-order sensitivity has little to read when the dependence is a 4-way product, whose first-order Fourier mass is near zero. An interaction-aware attribution would be the right instrument and we have not built one. As it stands, **QHA is a competitive high-order *predictor* but, at $k \geq 4$, a worse *locator* than the classical baselines** — a real limitation for this section's detector framing.

D.4 Supporting figures and tables

Fig. 2 plots Table 1 together with the k-junta control.

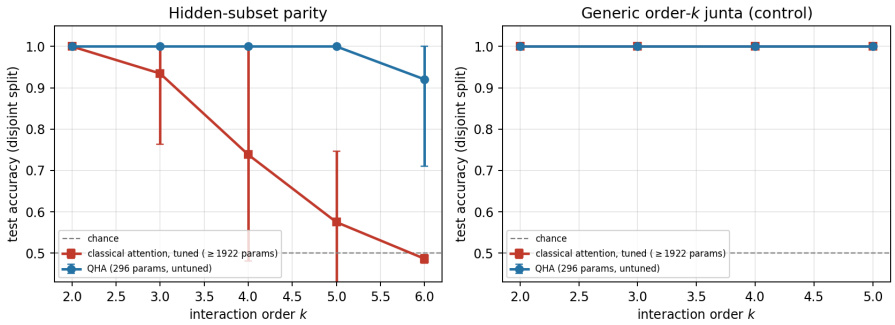


Figure 2: Test accuracy vs. order k (n=12, disjoint split; classical baseline tuned over 20 configurations, QHA untuned). **Left (parity):** the tuned head matches QHA through k=3, then falls to chance by k=6. **Right (k-junta control):** this target carries low-degree mass and the tuned head matches QHA at *every* k tested — the separation is specific to high-Fourier-degree structure, not to k.

Fig. 3 draws an explicit small QHA instance; Fig. 4 gives the order-k vs. qubit-count phase diagram behind Table 1; Fig. 5 plots the IBM Heron qubit scaling of Table 7; Fig. 6 shows the order-4 convergence traces; Fig. 7 collects the six ablations above; and Fig. 8 plots the epistasis efficiency frontier of Table 4. Table 5 reports causal support recovery, Table 6 the cross-domain confirmation (LPN and triangle detection), and Table 7 the per-n IBM Heron numbers.

Table 4: Higher-order epistasis — phenotype prediction (balanced accuracy, mean±std over 6 seeds, n=12, label noise 0.1; "p" = learned parameters, noise sweep in App. A). No main effects ⇒ linear/pairwise at chance for $k \geq 3$; MDR is the field-standard exhaustive search. QHA matches the MDR ceiling and leads the learned baselines, which lose more than QHA under fixed-final-epoch reporting. Last row: efficiency Eff $= (\text{acc} - 0.5)/\text{params} \times 10^3$ at k=4, scored only for models that reach the ceiling.

order	linear 26 p	pairwise 289 p	MDR exhaustive	HOFM 451 p	MLP 9.3k p	**QHA** **296 p**
k=2	0.50±.01	**0.90**±.00	0.90±.00	0.87±.01	0.84±.01	**0.90**±.00
k=3	0.50±.02	0.50±.01	0.90±.00	0.87±.01	0.83±.01	**0.90**±.00
k=4	0.50±.01	0.50±.01	0.90±.01	0.86±.01	0.83±.01	**0.90**±.00
Eff.$_{k=4}$	—	—	—	0.79	0.04	**1.34**

E Scope of the empirical claims

Scope. Hidden-subset parity and random k-juntas are the standard testbeds for high-order and compositional learning in learning theory and in the study of reasoning in sequence models (Bhattamishra et al., 2023; Abbe et al., 2022; Sanford et al., 2023); we use them because they isolate the one axis our theory concerns — interaction order. The hardware study is a faithfulness demonstration (simulation-trained parameters executed on IBM Heron for inference), showing the learned order-k rule survives device noise; it is not a hardware-training or speedup claim (Sec. 7).

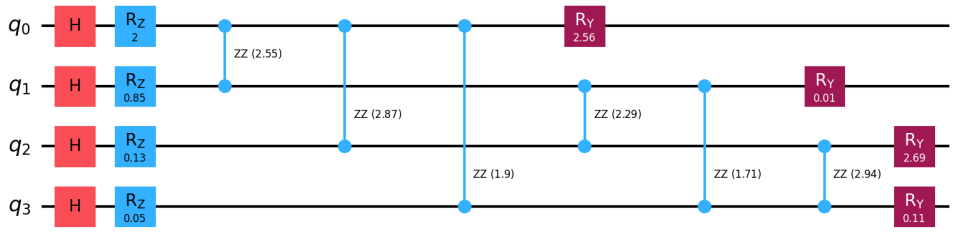


Figure 3: An explicit QHA circuit (n=4, L=1): Hadamards and R_Z data re-uploading, all $\binom{4}{2}$=6 R_{ZZ} couplings (all-to-all), a trainable R_Y layer, and a single-qubit $\langle Z \rangle$ read-out (not shown).

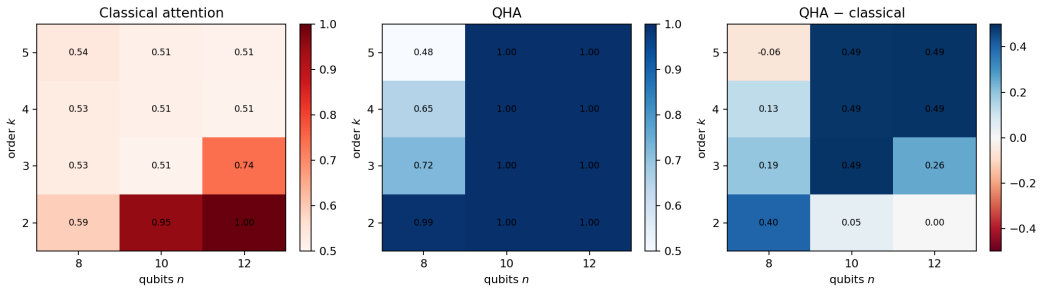


Figure 4: Order-k vs. qubit-count n phase diagram (parity, mean test accuracy) at the fixed 50-epoch budget for both models. QHA (middle) is at ceiling almost everywhere (n=8 is data-limited by the 256-point cube); the difference (right) localizes the advantage to the high-order region. Note that the classical panel (left) is the *untuned* baseline, which at k=3 understates it — see Table 1 for the tuned comparison at n=12.

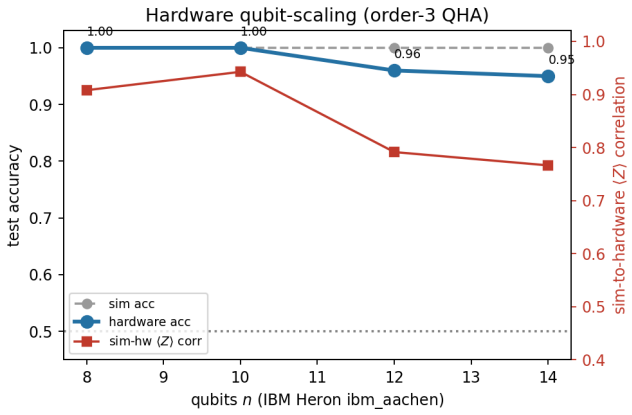


Figure 5: Hardware qubit-scaling on IBM Heron: hardware accuracy (blue) tracks the noiseless simulator (grey) up to n=10 and stays $\geq$0.95 through n=14, even as the raw $\langle Z \rangle$ correlation (red) decays — the linear read-out absorbs the noise.

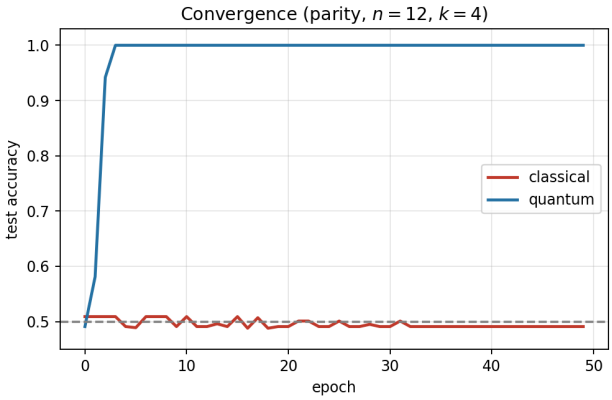


Figure 6: Convergence on order-4 parity (n=12) at the fixed 50-epoch budget: QHA reaches ceiling within a few epochs while the classical attention head stays at chance throughout. This is a statement about *this* budget, not about attention in general — given 100 epochs and a tuned learning rate the same head solves 4 of 8 supports at k=4 (Table 1). The contrast the figure makes is one of convergence speed.

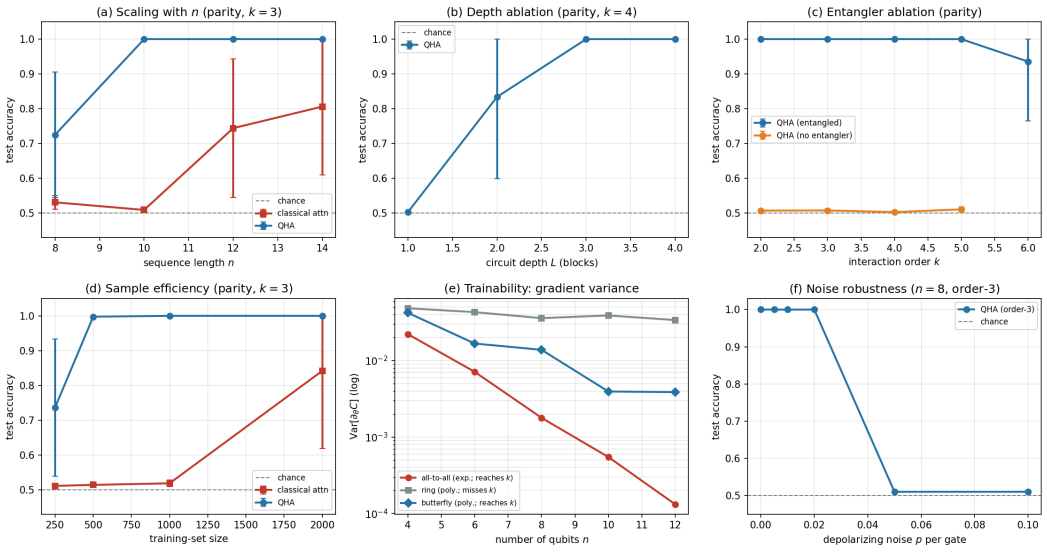


Figure 7: Extended studies. (a) scaling with n; (b) depth vs. reachable order; (c) entangler ablation; (d) sample efficiency; (e) gradient variance vs. qubits for three entanglers — all-to-all is exponential (barren) but reaches order-k; the ring is polynomial but misses order-k; the log-diameter butterfly is polynomial-fit *and* reaches order-k, narrowing the expressivity–trainability gap; (f) depolarizing-noise robustness. See Sec. 5.1.

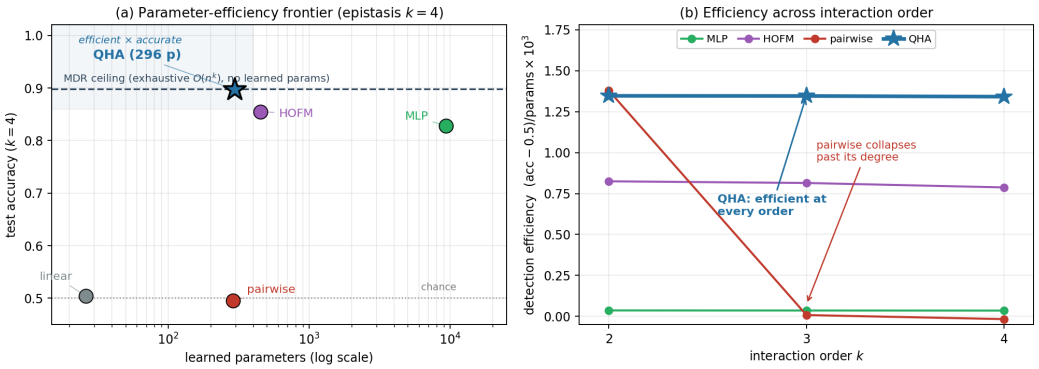


Figure 8: QHA as a compact high-order detector (epistasis, noise 0.1, 6 seeds, fixed final epoch). (a) **Parameter-efficiency frontier** at the hardest order k=4: QHA (star) sits in the "efficient × accurate" corner, reaching the MDR ceiling with the fewest learned parameters (296); under final-epoch reporting the learned baselines land slightly below that ceiling (HOFM 0.86 at 451 params, MLP 0.83 at 31× the parameters), MDR attains it only through exhaustive $O(n^k)$ search, and linear/pairwise sit at chance. (b) **Efficiency across interaction order**: $(\text{acc} - 0.5)/\text{params} \times 10^3$ vs. k. QHA stays efficient at *every* order (≈ 1.34); the pairwise model is competitive only at its own degree k=2, HOFM is uniformly lower (≈ 0.79), and the MLP sits near zero. (Linear and MDR omitted: a tiny-parameter artifact and no learnable parameters, respectively.)

Table 5: Causal support recovery on k-way epistasis (n=12, 3 seeds, noise 0.1): does the method identify the true interacting loci? Exact = recovered set equals the true k-subset; prec@k = fraction of true loci in the top-k. Neural/FM attributions use mean $|\partial(\text{logit gap})/\partial x_i|$. QHA recovers the support exactly through k=3; at k=4 first-order attribution degrades while prediction does not (see text).

	MDR	HOFM (attr.)	MLP (attr.)	**QHA (attr.)**
exact-recovery rate (k=3)	1.00	1.00	1.00	**1.00**
precision@k (k=3)	1.00	1.00	1.00	**1.00**
exact-recovery rate (k=4)	1.00	1.00	1.00	0.00
precision@k (k=4)	1.00	1.00	1.00	0.67

Table 6: Cross-domain confirmation (test accuracy, mean±std over 6 seeds, fixed final epoch; attention tuned over training length, as in Table 1). On LPN QHA is the most accurate model at the smallest budget (296 params); **on triangle detection it is not** — it ties the attention head (0.67) and trails HOFM (0.74). Last column is Eff $= (\text{acc} - 0.5)/\text{params} \times 10^3$ for QHA / HOFM / MLP. **Both use genuinely disjoint input splits** (LPN 2731/1365 of the 4096-point cube; triangle 140/70 of the 210 fixed-edge-count graphs); an earlier version did not, see App. A. On triangle detection *nothing works well* — the 10-qubit encoding admits only 210 graphs, so a disjoint split leaves 140 training examples — and Eff is not scored there because no model reaches the 0.9 threshold.

domain (task)	linear	attn	HOFM	MLP	**QHA**	Eff (QHA/HOFM/MLP)
Coding: LPN (k=4, 5% noise)	0.48±.01	0.70±.15	0.92±.01	0.90±.01	**0.95**±.00	**1.52** / 0.93 / 0.04
Graphs: triangle detection	0.62±.04	0.67±.05	**0.74**±.02	0.71±.05	0.67±.05	—/—/—

Table 7: IBM Heron (`ibm_aachen`) validation of trained order-3 QHA heads across qubit counts (100 inputs, 2048 shots). *sim* = noiseless statevector; *hw* = hardware; ρ = sim-to-hardware $\langle Z \rangle$ correlation.

qubits n	transpiled depth	sim acc	hw acc	ρ
8	465	1.00	1.00	0.91
10	435	1.00	1.00	0.94
12	590	1.00	0.96	0.79
14	999	1.00	0.95	0.77

F Detailed proofs

This appendix is written to be self-contained for an ML audience: each result is preceded by a short primer on the (non-ML) tool it uses. Throughout, a single classical self-attention layer is the model class $T^{1,H}_{1,m,1,p}$ of Sanford et al. (2023): one layer, H heads, embedding dimension m, p-bit fixed-point arithmetic, on a length-N input; we write $\text{budget} = mHp$ for its total "size".

Primer: Boolean functions, Fourier degree, and interaction order. Any function $f : \{-1,+1\}^n \to \mathbb{R}$ has a unique multilinear (Fourier) expansion $f(x) = \sum_{S \subseteq [n]} \hat{f}(S)\, \chi_S(x)$, where $\chi_S(x) = \prod_{i \in S} x_i$ is the parity (monomial) over the subset S and $\hat{f}(S) \in \mathbb{R}$ are the Fourier coefficients. The *degree* of f is the largest $|S|$ with $\hat{f}(S) \neq 0$; we call this the *interaction order* because χ_S couples exactly the $|S|$ tokens in S multiplicatively. Dot-product attention scores a pair (i,j) via $q_i^\top k_j$, a degree-2 form in the token embeddings; the layer's output is not itself degree-2 (the softmax is non-linear), but its pairwise scoring primitive is what makes reaching order $k > 2$ expensive, and the question is how expensive. The hidden-subset parity used in our experiments, χ_S with $|S| = k$, is the function with *all* Fourier mass on a single degree-k coefficient — the cleanest order-k target.

F.1 Lemma 1: order-k classical lower bound

Primer: communication complexity (the proof tool). In two-party communication complexity, Alice holds a and Bob holds b and they must compute $g(a,b)$ exchanging as few bits as possible; the minimum worst-case number of bits is the communication complexity of g. The canonical hard problem is set-disjointness DISJ_N: $a, b \subseteq [N]$ and $g = 1$ iff $a \cap b = \varnothing$; it requires $\Omega(N)$ bits (Razborov, 1992). In the r-party number-in-hand model Chakrabarti et al. (2003) prove $\Omega(N/(r \log r))$, improved to $\Omega(N/r)$ by Gronemeier (2009); either suffices for us, since $r = k - 1$ is a constant. The proof strategy is the standard "simulation" argument: (i) show a small attention layer induces a cheap communication protocol; (ii) show the target task requires expensive communication; (iii) conclude the layer cannot be small.

Proof of Lemma 1. **Step 1 (a small attention layer ⇒ a cheap protocol).** Partition the N input tokens among the players. Sanford et al. (2023) show that any $f \in T^{1,H}_{1,m,1,p}$ can then be evaluated by a protocol exchanging only $O(mHp \cdot \log\log N)$ bits: each head's softmax aggregation, at p-bit precision, is summarised by $O(m)$ numbers per token, so a player can transmit its head contributions in $O(mp)$ bits, and there are H heads. The bit count is thus controlled by the layer's $\text{budget} = mHp$, independent of which function is computed.

Step 2 (the order-k task is expensive). Consider the order-k matching predicate $\text{MATCH}_k(X)_i = \nexists\{\exists\, j_1, \ldots, j_{k-1} : x_i + \sum_t x_{j_t} \equiv 0 \pmod M\}$, the j_t ranging over distinct tokens other than i (for k=2 this is the pairwise "does some other token complete x_i to 0"; for general k it asks for a matching $(k-1)$-tuple). We reduce $(k-1)$-party set-disjointness DISJ^{k-1}_m to it.

A collision-free encoding. Write $p = k - 1$, let player r hold $S_r \subseteq [m]$, and put $B = 2m + 1$, $M = B^{2p}$. Player r contributes one token per $e \in S_r$, of value

$$\phi_r(e) \;=\; e\, B^{\,r-1} \;+\; (m - e)\, B^{\,r \bmod p} \;+\; B^{\,p+r-1} \pmod M,$$

and the query token is $x_i = -T$ with $T = \sum_{d=0}^{p-1} m\, B^d + \sum_{r=1}^{p} B^{\,p+r-1}$. Every coefficient in a sum of p such tokens lies in $[0, B)$: digits $p, \ldots, 2p-1$ count the tokens taken from each

block (at most $p < B$), and digits $0, \dots, p-1$ hold sums $e + (m - e') \in [0, 2m]$. **No carries occur, so base-B representations are unique and no spurious zero-sum tuple exists**: a p-tuple sums to T iff every digit matches, i.e. the tag digits force exactly one token per block and digit d reads $e_{d+1} + (m - e_d) = m$ (indices cyclic), forcing $e_1 = \cdots = e_p$. Hence some tuple gives $x_i + \sum_t x_{j_t} \equiv 0$ iff $\bigcap_r S_r \neq \varnothing$, with $M = (2m+1)^{2(k-1)} = N^{\Theta(1)}$ for fixed k. (For k=3 this specialises to the two-party reduction of Sanford et al., 2023, Thm 7.) Hence computing $\textsc{Match}_k$ solves DISJ_m^{k-1} with $m = \Theta(N)$.

Step 3 (combine). In the number-in-hand model DISJ_m^{k-1} needs $\Omega\big(m/((k-1)\log(k-1))\big)$ bits (Chakrabarti et al., 2003) — indeed $\Omega(m/(k-1))$ (Gronemeier, 2009). Matching this against the $O(mHp \log\log N)$ bits the layer affords (Step 1), with $m = \Theta(N)$, gives

$$\text{budget} = mHp = \Omega\Big(\tfrac{N}{k \log k \, \log\log N}\Big),$$

which for every *fixed* k is $\Omega(N/\log\log N)$. So no single attention layer with $mHp = o(N/\log\log N)$ computes $\textsc{Match}_k$ for $k \geq 3$; the $k = 3$ case recovers Sanford et al. (2023), Thm 7. □

Remark (honest scope). The multiparty disjointness bound we invoke degrades only polynomially in k (Chakrabarti et al., 2003; Gronemeier, 2009), so for the constant-k regime of our experiments the budget threshold is $\Omega(N/\log\log N)$ as stated; we do not claim anything uniform in growing k. The lemma is a *single-layer* statement; multi-layer lower bounds for the same task are conjectural (Sanford et al., 2023; 2024), and we do not claim them.

F.2 Theorem 1: the quantum construction

Primer: the Heisenberg picture (why we track observables). An expectation value can be computed by evolving the *state* forward through the gates, or equivalently by evolving the *measured observable* backward: for a unitary U, $\langle\psi|U^\dagger Z U|\psi\rangle$, so we may ask how $U^\dagger Z U$ looks. This "Heisenberg picture" is convenient because Clifford gates map a Pauli operator to another (signed) Pauli. In particular $\mathrm{CNOT}_{i\to r}$ (control i, target r) maps the target's $Z_r \mapsto Z_r Z_i$. Accumulating onto a read-out qubit r therefore turns $\langle Z_r\rangle$ into a *product* of input bits — a parity; choosing r *inside* the support (below) makes this product exactly χ_S, with no ancilla.

Proof of the quantum direction. We give an explicit circuit over the QHA gate set whose read-out $\langle Z_r\rangle$ on a support qubit $r \in S$ equals the order-k monomial $\chi_S(X) = \prod_{i\in S} x_i$ for any $|S| = k$. **The input state is the computational-basis encoding** $|X\rangle = \bigotimes_w |b_w\rangle$ with $b_w \in \{0,1\}$ and $x_w = (-1)^{b_w}$; this is an assumption on the input, not something the encoder of Eq. 1 produces — a diagonal R_Z cannot map $|0\rangle$ to $|1\rangle$, and we make no such claim (this is precisely the encoder gap discussed after Thm. 1).

What computes the parity. Choose the read-out qubit to be a *member of the support*, $r \in S$ (e.g. $r = \min S$), and apply $\mathrm{CNOT}_{i\to r}$ for each $i \in S \setminus \{r\}$; then measure Z_r. By the primer, after these $k-1$ CNOTs the observable has become $Z_r \mapsto Z_r \prod_{i\in S\setminus\{r\}} Z_i = \prod_{i\in S} Z_i$, so on a basis state $\langle Z_r\rangle = \prod_{i\in S} x_i = \chi_S(X)$, as claimed. Because r is itself in S, its own bit x_r supplies the remaining factor: *no ancilla is introduced* (the read-out is one of the n token qubits) and no spurious factor appears. Equivalently, qubit r accumulates the XOR (parity) of the support bits.

Why this lives in the QHA gate set. Each CNOT is realised from QHA's gates (H, R_Z, R_{ZZ}, R_Y) via

$$\mathrm{CNOT}_{i\to r} = (I \otimes H)\,\mathrm{CZ}_{i,r}\,(I \otimes H), \qquad \mathrm{CZ}_{i,r} = e^{i\pi/4}\, R_Z^{(i)}(-\tfrac{\pi}{2})\, R_Z^{(r)}(-\tfrac{\pi}{2})\, R_{ZZ}^{(i,r)}(\tfrac{\pi}{2}),$$

and the Hadamard is obtained from the per-qubit R_Y layer ($H = R_Y(\frac{\pi}{2})\, Z$, with the global phase and Pauli-Z absorbed into adjacent single-qubit rotations). Because the QHA entangler applies R_{ZZ} on *every* pair, all pairs (i, r), $i \in S \setminus \{r\}$, are available.

Precise scope: architecture class vs. trained encoder. We state exactly what is and is not proved. Theorem 1 is, by construction, a statement about the QHA *architecture class* (a gate-set witness), not about the re-uploading ansatz of Eqs. 1–3 that we train: n qubits (one per token, no ancilla),

the gate set $\{H, R_Z, R_{ZZ}, R_Y\}$, and a single-qubit $\langle Z_r \rangle$ read-out. The witness above is a concrete circuit *in this class* — every gate it uses (H, fixed-angle R_Z, and the CNOTs compiled from R_{ZZ} and single-qubit rotations as displayed) belongs to the set, at depth $O(\log k)$ with $O(k)$ two-qubit gates — whose $\langle Z_r \rangle$ *exactly* equals the Boolean monomial χ_S. It loads the binary inputs in the computational basis. The *trained* head injects inputs through the re-uploading rotation $R_Z(\theta^{\text{enc}}_{\ell w} x_w)$ rather than basis loading; we do *not* claim the two encoders are gate-for-gate identical. What is proved is *representability*: an order-k read-out is realizable within the QHA architecture class at logarithmic depth (Thm. 1) while no classical attention layer of budget $mHp = o(N/\log\log N)$ computes MATCH_k (Lemma 1). That the re-uploading instance *attains* this order is consistent with re-uploading raising the reachable Fourier degree to k (Schuld et al., 2021) and is what Sec. 5 confirms empirically — the role we assign to the experiments.

Resource count. Accumulating the parity with a balanced binary tree of CNOTs onto the support qubit r (rather than a chain) uses $O(k)$ two-qubit gates at circuit depth $O(\log k)$. Each tree level is a set of *disjoint* (commuting) R_{ZZ} couplings, which one QHA block's all-to-all entangler realizes in parallel; hence the construction needs only $O(\log k)$ QHA blocks (not $O(k)$). So a single QHA head *represents* every order-k monomial on a single-qubit read-out, with n qubits, depth $O(\log k)$ and $O(k)$ two-qubit gates. Together with Lemma 1, which shows that no single classical attention layer of budget $o(N/\log\log N)$ computes MATCH_k, these give the two complementary order-k results stated in Thm. 1; because the targets differ, they do *not* constitute a separation theorem. □

F.3 Remark 1: the fixed-budget capacity heuristic

Primer: why norm controls generalization. For bounded-norm models, the generalization gap (test minus train error) is controlled by a capacity measure (Rademacher complexity / covering number) that grows with the weight norm; schematically, gap $\lesssim \widetilde{O}(\text{norm}/\sqrt{m})$ for m samples. So if representing a target *requires* a large norm, a fixed-budget model cannot generalize it from limited data — it would need either more parameters/norm or far more samples.

What the classical literature does and does not give us. Edelman et al. (2022) are *constructive*: they exhibit bounded-norm Transformer(+MLP) parameter settings that represent sparse Boolean functions, and show that for symmetric sparse functions the required norm is polynomial while their construction for arbitrary sparse functions carries a 2^s factor in the FFN. Their covering-number bound (Thm. 4.2/Cor. 4.5) then gives Rademacher complexity $\widetilde{O}(\text{norm})$, hence gap $= \widetilde{O}(\text{norm}/\sqrt{m})$. **We must be explicit that this is an upper bound on what a construction costs, not a lower bound on what any head needs**: it does not prove that a fixed-budget attention head is *unable* to generalize an order-k rule. We therefore do not state a fixed-budget generalization *separation*. What we take from it is a capacity heuristic — the only known routes to representing a generic order-k rule inside a bounded-norm head are expensive in norm, and norm is what controls the generalization gap — which is why we test the fixed-budget, disjoint-split regime empirically rather than claim it by theorem. Establishing a genuine norm lower bound for attention on generic k-juntas is, to our knowledge, open, and we flag it as the missing piece of the theoretical picture.

Quantum side. Data re-uploading raises the reachable Fourier degree with circuit depth (Schuld et al., 2021); concretely, the witness circuit of Thm. 1 realizes the degree-k monomial at depth $O(\log k)$ with $O(k)$ two-qubit gates and $\text{poly}(n)$ parameters, with no $2^{\Omega(k)}$ norm blow-up. This is a representability statement; that a *trained* QHA head then generalizes the order-k rule at fixed budget is what Sec. 5 measures.

Fourier-degree refinement (matches the experiments). The classical bound bites in proportion to the target's *high-degree* Fourier mass. Parity puts all mass at degree k (worst case ⇒ maximal gap, Fig. 2 left). A balanced random k-junta spreads mass over degrees $\leq k$; its low-degree part is learnable by the low-degree-biased head (Bhattamishra et al., 2023), so the classical model keeps up until k is large (Fig. 2 right). This is precisely the empirical trend.

F.4 THEOREM 2: ABSENCE OF BARREN PLATEAUS

Primer: what a barren plateau is. A variational circuit has a *barren plateau* if the cost gradient's variance vanishes exponentially in the number of qubits, $\mathrm{Var}_\theta[\partial_\mu C] = 2^{-\Omega(n)}$; then gradients are exponentially small almost everywhere and training is infeasible at scale. Cerezo et al. (2021) prove this is *avoided* when two conditions hold: the circuit is *shallow* (depth $O(\log n)$) and the cost is built from *local* observables (acting on $O(1)$ qubits) rather than a global one. QHA's read-out $\langle Z_w \rangle$ is 1-local by construction, which is the key enabling property.

Proof of Thm. 2. Take the QHA variant whose entangling layers form an alternating block ansatz of local 2-designs on bounded-range qubits, at depth $L = O(\log n)$, with the 1-local cost $C(\theta) = \langle Z_w \rangle$. These are exactly the hypotheses of Cerezo et al. (2021), Thm 2, whose conclusion is an inverse-polynomial lower bound on the gradient variance,

$$\mathrm{Var}_\theta[\partial_\mu C] \;\geq\; \Omega\big(1/\mathrm{poly}(n)\big),$$

so there is no barren plateau: gradients shrink only polynomially, and the head is trainable at scale. The locality of the read-out is what places the cost in the regime where this lower bound (rather than the exponentially small global-cost bound) applies. □

Scope (stated honestly). Thm. 2 covers the shallow local-2-design variant, to which Cerezo et al. (2021) applies rigorously. The all-to-all R_{ZZ} entangler used in the experiments is more expressive and is *not* a geometrically local 2-design, so the theorem does not directly certify it. Its measured gradient variance in fact decays *exponentially* ($\mathrm{Var} \sim 2^{-0.9n}$, Fig. 7e) — a barren plateau in the asymptotic sense — yet it remains trainable at the tested sizes ($n \leq 14$, all $k \leq 6$; order-2 parity is fit within 4 epochs) despite that decay, because the gradient signal is still resolvable there. The log-diameter butterfly entangler is the variant that reaches the order-k ceiling *and* shows a polynomial-fit decay (Sec. 5.1e), and is what we would scale up. As Sec. 7 discusses, the same locality/shallowness that guarantees trainability also tends to imply classical simulability (Cerezo et al., 2025); we therefore present trainability as a *theorem for the benign variant* and an *empirical observation for the expressive one*, and make no exponential-speedup claim.

AI USE STATEMENT

Parts of this paper's text, code and analysis were produced with the assistance of generative AI. The authors take full responsibility for the final content of this paper, including any text produced with the assistance of generative AI.