MAGICCORE-LM: COMPACT ALGEBRAIC INTERFACES FOR LEARNING NEAR-CLIFFORD QUANTUM CIRCUITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly applied to specialized scientific problems, often using textual or symbolic descriptions as inputs. However, their general-purpose capabilities do not automatically translate into an understanding of the domain-specific mathematical structure underlying these problems. A faithful description may therefore specify what to compute without exposing the structure needed to learn it effectively. We address this gap in quantum computing, where different gate sequences can implement the same computation while obscuring the algebraic relations that determine measurement outcomes. We introduce MAGICCORE-LM, a learning interface that uses quantum algebra to translate near-Clifford circuits into compact, unexecuted rotation instructions and transformed measurement queries. The compiler performs exact Clifford transformations, while a relation-aware reader learns how the resulting operation identities, signs, and physical order jointly determine measurement expectations. In a three-seed blind evaluation of eight matched variants, MagicCore reduces in-distribution (IID) mean absolute error by 12.7% relative to raw-gate inputs. Standard set-based readers also achieve lower IID error with Compact inputs, supporting the interface across reader architectures. Deploying the frozen reader with an optimized GPU compiler yields $1.68\times$ and $9.15\times$ end-to-end speedups over Raw on the audited higher-complexity circuits at batch sizes 1 and 32, respectively, on an RTX 4090. These measurements include preprocessing; compiler changes preserve same-batch predictions without retraining. Additional studies examine frame-explicit and propagated-state representations, as well as integration with five pretrained LLMs.

1 INTRODUCTION

Language models increasingly support tasks specified through text (Brown et al., 2020), including scientific workflows (Boiko et al., 2023). Yet model capability and problem representation are different design choices: structured inputs supply useful inductive biases (Battaglia et al., 2018). A sequence can describe a calculation without exposing its mathematical structure. We study the interface question: *which known structure should be computed exactly, and how should the remaining problem be presented for learning?*

Quantum expectation prediction provides a test: given a circuit and query, predict the measurement's mean. Figure 1 expands 36 gates to 1,028 without changing the computation. A raw-gate learner must accommodate this variation while learning the numerical map. The challenge is therefore not merely to shorten an input, but to separate changes in its description from changes in the calculation.

Near-Clifford circuits offer an exact mathematical foothold. Clifford operations admit compact classical descriptions (Gottesman, 1998; Dehaene & De Moor, 2003; Aaronson & Gottesman, 2004); non-Clifford gates supply the *magic* resource needed for universal computation (Howard et al., 2014; Veitch et al., 2014; Howard & Campbell, 2017). Neutral Clifford rewrites increase raw length; non-Clifford rotations can increase algebraic work. Exact algebra separates these changes while preserving the queried computation.

Circuit-image CNNs (Cantori et al., 2023) and circuit-DAG GNNs (Liu et al., 2025) learn from the gate scaffold; circuit distribution (Cantori & Pilati, 2025) and non-Clifford complexity (Du et al., 2025) also shape predictability. Our research question is: *can known circuit algebra be organized*

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.

into unpropagated instructions that are easier to learn from than raw gates, under matched training and model budgets? Crucially, a useful translation should remove irrelevant syntax while leaving a genuine numerical task to the learner.

MAGICCORE-LM is an *algebra-to-model translator*. Clifford/Pauli algebra changes coordinates exactly. Our Compact interface specifies what crosses the compiler–learner boundary: typed, signed rotation instructions and queries, with physical order distinguished from storage position. The compiler leaves these rotations unexecuted; a reader learns their joint numerical effect under a shared query constraint. Our design contribution is this explicit handoff and its structured reader (Figure 2; background in Section 2).

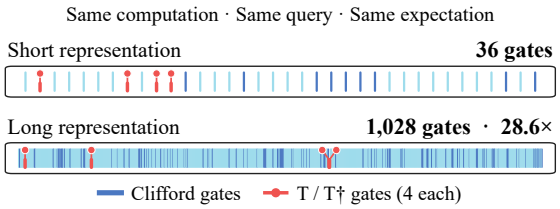


Figure 1: **Same computation, different raw syntax.** Clifford identity rewrites expand 36 to 1,028 gates (28.6×), with U, Q, $t = 4$, and $\langle Q \rangle = \sqrt{2}/4$ unchanged. Horizontal positions are normalized separately; pin leaders separate adjacent T gates.

Two linked tests organize the evidence. *Does Compact improve learning?* Table 1 compares eight equally budgeted input–reader combinations, frozen across three seeds before blind evaluation. MagicCore leads mean IID/OOD MAE and reduces IID Raw error by 12.7%; ordinary set readers also improve IID error. *Can this representation be deployed efficiently?* Table 3 deploys the frozen MagicCore reader: exact compiler reimplementations preserve its inputs and same-batch predictions while lowering measured end-to-end latency relative to Raw, including the front end. Propagated-state and sidecar–LLM studies are extensions with separate computational boundaries.

This suggests a route to scientific model reuse: encode exact domain knowledge in an interface shared by independently replaceable compilers and readers. Compiler improvements then preserve trained models, while reader comparisons preserve the mathematical input. Our contributions are:

- **A learning-oriented Compact contract.** Unpropagated instructions retain Pauli support, signs, physical order, and queries independently of neutral Clifford length.
- **Structured numerical prediction.** Relation-aware reading learns interactions between instructions; a shared query-linear head imposes a numerical constraint across reader controls.
- **Learning-to-deployment evidence.** A three-seed fresh-blind comparison tests interface and reader choices; a six-path compiler audit preserves the selected models' inputs while measuring complete cost and no-retraining deployment savings.

2 PRELIMINARIES

From bits to quantum circuits. A classical bit is either 0 or 1. A *qubit* has state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, where complex amplitudes satisfy $|\alpha|^2 + |\beta|^2 = 1$. Measurement in this basis returns 0 or 1 with probabilities $|\alpha|^2$ and $|\beta|^2$; relative phases affect later interference. A circuit applies reversible linear transformations called *unitary gates*; their ordered product U maps the initial state $|0^n\rangle$ to $U|0^n\rangle$.

The prediction query. The Pauli operators I, X, Y, Z describe identity and three elementary spin measurements. A *Pauli string* is their tensor product: $Q = Z_1 X_3$, for example, measures qubits 1 and 3 along the Z and X axes and is identity elsewhere. For a non-identity string, each measurement yields $+1$ or -1; we predict its mean, $y(U, Q) = \langle 0^n | U^\dagger Q U | 0^n \rangle \in [-1, 1]$, not the full quantum state. This expectation is a real-valued target, and different queries on one circuit can have different answers. Instead of evolving the state, one can equivalently propagate the query backward as $Q \mapsto U^\dagger Q U$ and evaluate it on $|0^n\rangle$.

Easy structure and branching structure. Our gate set is $H, S, S^\dagger, \mathrm{CX}, T, T^\dagger$; a dagger denotes an inverse gate. The first four are *Clifford* gates: conjugating a Pauli yields another signed Pauli, e.g., $HZH = X$ (Gottesman, 1998; Dehaene & De Moor, 2003; Aaronson & Gottesman, 2004). Thus an entire Clifford circuit is specified, up to an irrelevant global phase, by the signed images of the $2n$ generators $X_1, \ldots, X_n, Z_1, \ldots, Z_n$, called its *frame*. The non-Clifford gates $T/T^\dagger$ supply

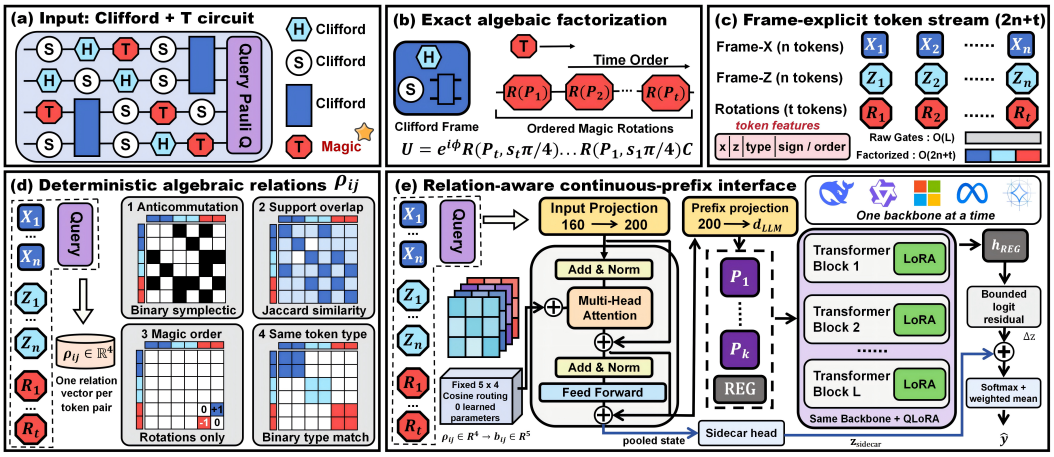


Figure 2: **MagicCore's frame-explicit interface and LLM extension.** **(a)** Circuit and query. **(b)** Exact Clifford factorization, up to global phase, into a frame and ordered rotations. **(c)** $2n{+}t$ typed tokens: n frame-X, n frame-Z, and t rotations. **(d)** Four relations: anticommutation, support overlap, signed order, and token type. **(e)** A frozen sidecar supplies continuous prefixes (Li & Liang, 2021) and base logits; the LoRA-adapted LLM (Hu et al., 2021) adds a bounded logit residual for posterior-mean prediction. Table 1 instead uses $t + 2$ unpropagated Compact records: the frame is absorbed into rotations and two queries, with phase relations and no LLM. Section 4 distinguishes these contracts from propagated branches.

magic (Howard et al., 2014; Veitch et al., 2014; Howard & Campbell, 2017): up to global phase, they are $R(Z, \pm\pi/4)$, where $R(P, \theta) = e^{-i\theta P/2}$. Unlike a Clifford gate, a non-Clifford rotation can turn one Pauli query into a weighted sum of Paulis. We call these terms *branches*.

The identity $CR(P, \theta) = R(CPC^\dagger, \theta)C$ collects Clifford operations into a frame while updating rotation axes. Order matters: Paulis can anticommute ($PQ = -QP$), and their rotations need not commute. This separates raw gate length L from the number t of non-Clifford rotations, active support k (the qubits used by the reduced problem), and the branch count produced during propagation. Identity rewrites increase L without changing the computation; larger t or k can increase the remaining algebraic work. Compact inputs retain the instructions; propagated inputs expose the branches produced by propagating the query through them.

3 Related Work

Quantum structure and prediction. Near-Clifford simulation organizes cost around non-Clifford count, stabilizer rank/extent, or quasiprobability negativity rather than raw gate count (Pashayan et al., 2015; Bravyi & Gosset, 2016; Bravyi et al., 2016; 2019). We organize these algebraic tools into learning inputs. Related predictors consume circuit images (Cantori et al., 2023; Cantori & Pilati, 2025), DAGs (Liu et al., 2025), or graphs of reliability/architecture properties (Wang et al., 2022); learning linear properties of bounded-gate circuits is also studied theoretically (Du et al., 2025). We adapt and retrain the closest architecture families on our shared task. Classical shadows use state measurements (Huang et al., 2020), and POVM-based Transformers simulate probability-state dynamics (Carrasquilla et al., 2021); we instead learn a reusable circuit/query-to-expectation map without measurement inputs.

Structured neural interfaces. Transformers (Vaswani et al., 2017), relative attention (Shaw et al., 2018), GAT (Veličković et al., 2017), Graphormer (Ying et al., 2021), Deep Sets (Zaheer et al., 2017), and Set Transformer (Lee et al., 2019) supply tools for structured reading. Symmetry-informed learning motivates task-aligned bias (Meyer et al., 2023). We hold the compact schema fixed across readers and separate positions from relations. Adapters (Houlsby et al., 2019), prefixes (Li & Liang, 2021), LoRA (Hu et al., 2021), and QLoRA (Dettmers et al., 2023) adapt parameters; our complementary choice is the state interface and numerical supervision.

Table 1: **Compact learning-interface confirmation.** Three new seeds; mean ± sample SD on fresh blind data. P: absolute position; R: algebraic relations (✓: on; ×: off); $^{\mathrm{T}}$: Transformer. **Best** and second-best means.

				Fresh IID	Fresh OOD	
Input / reader	**P**	**R**	**Params**	MAE ↓	MAE ↓	R^2 ↑
Raw$^{\mathrm{T}}$	✓	×	1.963M	0.2344 ± 0.0116	0.2377 ± 0.0092	-0.098 ± 0.151
Compiler-normalized Raw$^{\mathrm{T}}$	✓	×	1.963M	0.2397 ± 0.0133	0.2431 ± 0.0094	-0.141 ± 0.147
Compact$^{\mathrm{T}}$	✓	×	1.963M	0.2237 ± 0.0063	0.2434 ± 0.0138	-0.184 ± 0.217
Compact$^{\mathrm{T}}$	×	×	1.963M	0.2150 ± 0.0081	0.2298 ± 0.0046	-0.023 ± 0.049
Compact$^{\mathrm{T}}$	✓	✓	1.963M	0.2097 ± 0.0022	0.2342 ± 0.0051	-0.108 ± 0.062
MagicCore$^{\mathrm{T}}$	×	✓	1.963M	$\mathbf{0.2046 \pm 0.0039}$	$\mathbf{0.2272 \pm 0.0019}$	-0.084 ± 0.026
Compact DeepSets	×	×	1.973M	0.2186 ± 0.0010	0.2315 ± 0.0010	$\mathbf{-0.011 \pm 0.009}$
Compact Set Transformer	×	×	1.963M	0.2171 ± 0.0001	0.2322 ± 0.0021	-0.025 ± 0.025

4 MagicCore Learning Interface

MagicCore separates exact coordinate changes, Compact record organization, and a learned instruction-to-response map, allowing compiler optimization without changing reader inputs.

4.1 Exact compilation: established algebra

For $U = G_L \cdots G_1$ and a non-identity Pauli Q with canonical sign $+1$, we predict

$$y(U, Q) = \langle 0^n | U^\dagger Q U | 0^n \rangle. \tag{1}$$

Standard binary-symplectic updates collect Clifford gates into a frame (Dehaene & De Moor, 2003; Aaronson & Gottesman, 2004; Bravyi et al., 2016):

$$U = R(P_t, s_t \pi/4) \cdots R(P_1, s_1 \pi/4) C, \qquad R(P, \theta) = e^{-i\theta P/2}. \tag{2}$$

Here C is Clifford, P_j are conjugated Paulis, and $s_j \in \{-1, +1\}$ absorbs negative-axis signs. This exact coordinate change preserves computation up to global phase. Raw/factorized propagation agrees on generated examples, with independent state-vector checks on a fixed subset.

4.2 Interface design: compact state organization

For the anticommuting queries Q_1, Q_2 of Section 4.4, absorb the frame in $U = MC$: $P'_j = C^\dagger P_j C$, $Q'_a = C^\dagger Q_a C$, and $M' = C^\dagger M C$. Then $y(U, Q_a) = \langle 0^n | M'^\dagger Q'_a M' | 0^n \rangle$. Only Clifford conjugation is executed; the t rotations remain unexecuted instructions. Pauli axes, signs, physical indices, and two query identities form $t + 2$ records. Rotation-axis signs enter s_j; query signs enter observable weights. No terminal branch, compatibility flag, or component expectation is supplied.

Inactive qubit columns are removed and active columns canonically ordered by rotation/query signatures, preserving expectations on $|0^n\rangle$. Physical indices stay attached to records: reordering storage does not reorder the circuit. All readers receive the same algebraic problem.

The frame-explicit variant (Figure 2) lists $2n$ signed generator images and t rotations, plus a query. Both record counts are independent of neutral L at fixed n, t, with record width proportional to support (Appendix A).

4.3 Structured reading: fixed relations, learned interactions

Four local relations encode symplectic parity, real/imaginary Pauli-product phase, and signed rotation-order difference, not propagated coefficients. For dot-product score a_{ijh}, relation-aware attention

Table 2: **Frame-explicit learning under neutral syntax changes.** Mean ± sample SD over seven matched seeds: prospective seed 3407 and six post-hoc replications. The mean OOD advantage is $\Delta^* = 0.0603$ (hierarchical 95% CI $[0.0200, 0.1044]$). BMAE averages MAE over target classes; OOD also averages rewrite depths. Exact-5: five-class argmax accuracy. CNN/GNN baselines are task-adapted and retrained. **Best**; second best.

Method	Balanced MAE ↓ IID	Balanced MAE ↓ OOD	Held-out comparison: Gap to ours	Held-out comparison: Exact-5 Acc. (%) ↑	Model size: Params
Task-adapted published architectures					
Circuit-image CNN (Cantori et al., 2023)	0.4244 ± 0.0001	0.4243 ± 0.0002	+0.0734 ± 0.0590	26.1 ± 1.2	1.87M
Circuit-DAG GNN (Liu et al., 2025)	0.4243 ± 0.0000	0.4243 ± 0.0000	+0.0733 ± 0.0590	24.7 ± 5.1	1.94M
Matched internal architectures					
Raw-gate Transformer	0.4243 ± 0.0000	0.4243 ± 0.0001	+0.0733 ± 0.0590	21.7 ± 1.3	1.96M
Raw-gate Relation Transformer	0.4244 ± 0.0002	0.4243 ± 0.0000	+0.0733 ± 0.0590	22.3 ± 1.4	1.96M
Factorized-token Transformer	0.4130 ± 0.0086	0.4112 ± 0.0109	+0.0603 ± 0.0619	37.7 ± 3.0	1.96M
MAGICCORE-LM	**0.3536 ± 0.0551**	**0.3509 ± 0.0590**	–	**50.5 ± 8.5**	1.96M

(Vaswani et al., 2017; Shaw et al., 2018; Ying et al., 2021) gives

$$\widetilde{a}_{ijh} = a_{ijh} + \sum_{r=1}^{4} D_{hr} \tanh\left(\rho_{ij}^{(r)}\right), \tag{3}$$

where fixed discrete-cosine matrix $D \in \mathbb{R}^{5\times 4}$ routes relations to heads without learned parameters (Appendix B.1). Input projections, attention/MLP updates, and heads are trained. MagicCore retains physical order but omits absolute positions, preserving prediction under joint permutations of rotation records and relation axes. The factorial control and set readers share these compact records.

4.4 Learned mapping and numerical supervision

The continuous task queries $O_\alpha = \cos\alpha\, Q_1 + \sin\alpha\, Q_2$, where $Q_1Q_2 = -Q_2Q_1$, with target $y = y(U, O_\alpha)$. Each reader produces $u_\theta \in \mathbb{R}^2$ and two responses in one forward pass:

$$r_\theta = \frac{u_\theta}{\sqrt{1 + \|u_\theta\|_2^2}}, \qquad \hat{y} = w^\top r_\theta, \qquad \mathcal{L} = (\hat{y} - y)^2. \tag{4}$$

Anticommutation bounds the response pair to the unit disk. Known weights w contain $\cos\alpha, \sin\alpha$, adjusted for query signs, and enter only the final readout. All arms learn two query responses through this shared head and joint-target MSE, without component supervision. TRAIN-only augmentation varies observable mixing angle α, not circuit rotation angles.

Discrete tasks couple classification and expectation MSE; propagated continuous tasks regress a joint scalar (Appendix A).

4.5 Extensions beyond the Compact contract

Propagated-state boundary. This extension executes query propagation through rotations and the frame before learning, testing computed-state reading rather than the instruction map in Table 1. Final Pauli bits and signed coefficients c_b permit analytic contraction: $\langle Q \rangle = \sum_b c_b \mathbf{1}[x_b = 0]$ on $|0^n\rangle$, whether the compatibility flag is explicit or reconstructed. Anticommuting rotations can double terms: at most 2^t per query before merging. Propagation is fixed; the reader is trained (Appendix B.4).

Local-LLM integration. An auxiliary reader (*sidecar*) is trained first and frozen. Projected sidecar states form continuous prefixes for a LoRA-adapted backbone; a regression token supplies a bounded

correction:

$$z = z_{\text{sidecar}} + b \tanh\big(g_\theta(h_{\text{REG}})\big). \tag{5}$$

Here z and z_{sidecar} are class-logit vectors; the posterior-weighted class centers give the scalar prediction. Routes share b and the adaptation budget within each backbone (Appendix A.6). The Compact cost audit instead deploys instruction readers without an LLM.

5 Experimental Setup

For Compact learning, the eight arms in Table 1 share 3,840 training and 200 validation cores at $t \in \{5, 7\}$ with four active qubits, 7,200 updates, and a query-linear head. Equal-budget searches fix per-method optimizers; data/augmentation schedules and validation-only selection are matched. All 24 runs across three new seeds are frozen before generating 400 fresh IID cores at $t \in \{5, 7\}$ and 800 fresh OOD cores at $t \in \{9, 11, 13, 15\}$. Normalized Raw resynthesizes Clifford blocks, retaining $T/T^\dagger$ boundaries.

The primary comparison, prospectively specified and hash-frozen before blind generation, is fresh-OOD MAE of position-free compact vanilla minus MagicCore. Other comparisons are secondary; 95% paired intervals resample seeds and independent cores within t strata, without multiplicity adjustment. MAE, R^2, zero/train-mean references, the amendment, and all outcomes are in Appendix B.1. Deployment reuses the selected MagicCore and both Raw controls on shared cores (Table 3; Appendix B.2).

Supporting frame-explicit studies include retrained circuit-image CNNs (Cantori et al., 2023) and circuit-DAG GNNs (Liu et al., 2025). Propagated continuous/family and sidecar–LLM studies use separate schemas and cost boundaries; Appendix A specifies their budgets and release protocols.

6 Results

6.1 Compact Learning: Interface and Reader Controls

Table 1 tests input and reader choices without propagated branches. MagicCore has the lowest mean IID/OOD MAE, $0.2046 \pm 0.0039/0.2272 \pm 0.0019$. Its IID improvement over Raw is 0.0298 (12.7%), with all three paired seeds improving; the secondary 95% interval is $[0.0132, 0.0500]$. The Raw-relative OOD interval crosses zero. Against compiler-normalized Raw, MagicCore's gaps are 0.0351 IID and 0.0159 OOD, both with positive secondary intervals (Appendix Table 14).

Interface value across readers. Position-free compact vanilla, DeepSets, and Set Transformer improve IID MAE over both Raw baselines in all three seeds, with positive pointwise 95% intervals in direct post-hoc comparisons (Appendix Table 16). Their OOD intervals versus Raw cross zero, with gains in 2/3, 3/3, and 1/3 seeds, respectively (Table 17). MagicCore's primary OOD gain over compact vanilla is 0.00254 (1.11%; 95% CI $[-0.00594, 0.01224]$), with two of three seeds improving.

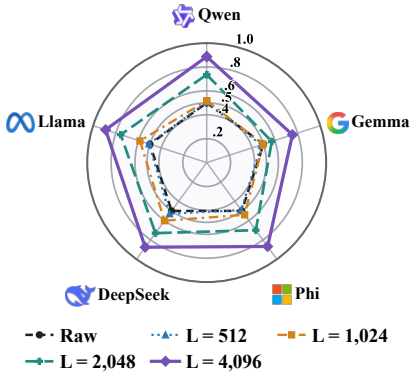


Figure 3: **Neural-path scaling across backbones.** Black: Raw parity (0.5); colors: MagicCore at the four lengths in the legend. Scores $1/(1 + c/c_{\text{Raw}})$ increase outward as latency falls. Costs: Appendix Tables 9–12.

6.2 Compact Deployment: Complete Cost Without Retraining

Does the learning benefit survive the cost of constructing Compact inputs? Table 3 reuses the selected MagicCore reader and both Raw controls. Four Compact compilers preserve input tensors on all 1,200 cores and same-batch predictions, isolating deployment from model changes. OOD median single-query end-to-end latency (E2E) is 6.98 ms with packed CPU or 5.18 ms with fused GPU versus Raw's 9.33 ms. Paired geometric-mean Raw-relative GPU speedups are $1.68\times$ (95% CI $[1.48, 1.88]$) at batch 1 and $9.15\times$ ($[7.76, 11.60]$) at batch 32. E2E includes compilation, encoding, transfers, and

prediction, with warm runtime and uncached inputs. On these workloads, Compact pairs lower mean error with lower complete latency, without retraining. Appendix B.2 retains six paths, shared 32-core timing cohorts per split, stage costs, and memory; batch-32 times cover whole batches. Across the full test sets, mean input lengths fall from 200/206 Raw records to 8/14 Compact records (IID/OOD). The compiler replaces gate syntax with algebraic instructions, shortening the neural sequence without truncating the underlying circuit.

Table 3: **Frozen Compact deployment cost.** Three seeds, no retraining; MAE: full-data mean ± SD. E2E medians include the front end; at $B = 32$, latency covers the entire batch. **Best** and second-best displayed values. All six paths: Appendix Table 18.

	MAE ↓		IID E2E (ms) ↓		OOD E2E (ms) ↓	
Input / compiler	IID	OOD	$B = 1$	$B = 32$	$B = 1$	$B = 32$
Raw	0.2344 ± 0.0116	0.2377 ± 0.0092	8.28	126.59	9.33	135.06
Normalized Raw	0.2397 ± 0.0133	0.2431 ± 0.0094	11.61	261.27	12.88	302.49
Compact / packed CPU	$\mathbf{0.2046 \pm 0.0039}$	$\mathbf{0.2272 \pm 0.0019}$	**6.33**	40.18	6.98	48.40
Compact / fused GPU	$\mathbf{0.2046 \pm 0.0039}$	$\mathbf{0.2272 \pm 0.0019}$	6.47	**15.78**	**5.18**	**16.04**

6.3 Extensions: other state contracts and LLM pipelines

These extensions use distinct state contracts (Section 4.5).

Frame-explicit replication. Seed 3407 and six post-hoc replications yield six wins and an OOD-BMAE gap of 0.0603 (95% CI $[0.0200, 0.1044]$; Table 2). Relations/positions vary jointly; byte-identical factorized OOD inputs test neutral-syntax robustness.

Computed-state complexity transfer. The branch-state/sidecar/Qwen pipeline reaches OOD BMAE 0.0408 ± 0.0126 versus 0.2601 for the near-constant Raw/Factorized predictions (zero reference: 0.26011), lower at every t (Figure 4; Appendix A.6). The Raw-relative gain is 0.2194 (CI $[0.2097, 0.2332]$); Appendix Table 13 gives per-t results and accuracy.

Propagated continuous prediction. On 600 fresh joint-observable targets outside the legacy centers, single-forward MagicCore reaches MAE 0.1773 ± 0.0129, $R^2 = 0.347 \pm 0.076$ (Table 4). The Raw-relative gain is 0.0620 (CI $[0.0467, 0.0763]$), positive for every seed and unseen complexity.

Propagated-state third-party transfer. The frozen pipeline transfers to 600 Qiskit `QAOAAnsatz/iqp` (Javadi-Abhari et al., 2024; Farhi et al., 2014; Bremner et al., 2011) and Qiskit Nature UCCSD/Jordan–Wigner (Qiskit Nature developers and contributors, 2023; Romero et al., 2019) circuits, independently labeled by PennyLane `default.qubit` (Bergholm et al., 2018). From synthetic $t \in \{5, 7\}$ training to 8/16/12-rotation tests, MAE is 0.1863 ± 0.0206 (Table 5).

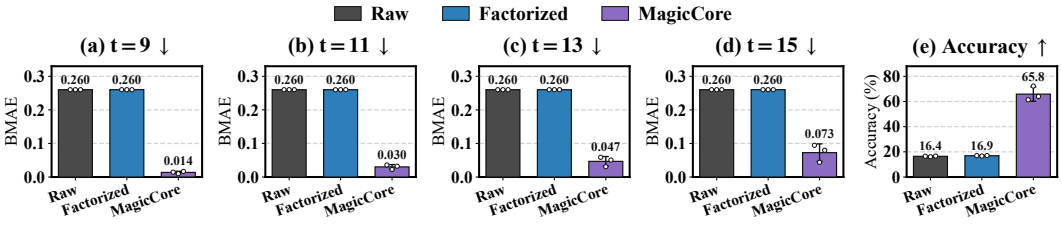


Figure 4: **Qwen3-4B extrapolation beyond the observed non-Clifford complexity.** Train at $t \in \{5, 7\}$, freeze, then test fresh $t \in \{9, 11, 13, 15\}$ circuits. Bars: three-seed means; error bars: sample SD; white points: individual seeds. The comparison includes MagicCore branch propagation and sidecar/Qwen fusion. Exact values and paired CIs: Appendix Table 13.

Table 4: **Single-forward extrapolation to genuinely continuous observables.** Matched 2M-parameter Transformers, frozen before 600 fresh cores; contractible branch inputs (Section 4). Mean ± sample SD over three seeds; MAE ↓, R^2/sign accuracy ↑. **Best**; second best. Training and paired gains: Appendix A.

Interface	$t = 9$ MAE	$t = 11$ MAE	$t = 13$ MAE	$t = 15$ MAE	Average MAE	R^2	Sign (%)
Raw	$.2318 \pm .0012$	$.2279 \pm .0032$	$.2527 \pm .0016$	$.2446 \pm .0016$	$.2392 \pm .0006$	$-.060 \pm .009$	50.4 ± 1.4
Factorized	$.2537 \pm .0071$	$.2417 \pm .0077$	$.2709 \pm .0168$	$.2507 \pm .0181$	$.2542 \pm .0094$	$-.266 \pm .078$	48.9 ± 2.2
MAGICCORE-LM	$.1397 \pm .0217$	$.1660 \pm .0115$	$.1938 \pm .0101$	$.2096 \pm .0089$	$.1773 \pm .0129$	$.347 \pm .076$	74.2 ± 4.2

Table 5: **Third-party fresh-blind transfer to standard algorithms.** No retraining; 600 Qiskit/Qiskit Nature circuits (200 per family), independently labeled by PennyLane. Three-seed mean ± sample SD. **Best**; second best. Paired gains: Section 6.

Interface	Qiskit QAOA MAE	Qiskit IQP MAE	Nature UCCSD MAE	Overall MAE	R^2	Sign (%)
Raw	$.4507 \pm .0149$	$.3839 \pm .0507$	$.7044 \pm .0111$	$.5130 \pm .0154$	$-.475 \pm .108$	50.1 ± 2.4
Factorized	$.4093 \pm .0183$	$.3281 \pm .0046$	$.7071 \pm .0094$	$.4815 \pm .0062$	$-.254 \pm .039$	49.4 ± 1.4
MAGICCORE-LM	$.1126 \pm .0187$	$.0775 \pm .0083$	$.3689 \pm .0355$	$.1863 \pm .0206$	$.745 \pm .050$	94.2 ± 1.5

Paired Raw/Factorized gains are 0.3267 ($[0.2904, 0.3692]$)/0.2952 ($[0.2707, 0.3202]$), positive for every fixed-angle family. Appendix Table 25 reports synthetic-family transfer.

Independent cores. On cores excluded from fitting Table 2 models, the seven-seed frame-explicit replication yields an OOD gap of 0.0564 (hierarchical 95% CI $[0.0198, 0.0978]$; Table 6).

Table 6: **Seven-seed component replication.** Fixed independent held-out set; mean ± sample SD. Seed 3407 is prospective; six seeds are post hoc. †: matched Transformer; Exact-5: five-class argmax accuracy. **Best**; second best.

	Components		BMAE ↓		OOD Exact-5 ↑
Method	**Fact.**	**Rel.**	**IID**	**OOD**	**Acc. (%)**
Raw-gate†	–	–	0.4243 ± 0.0001	0.4243 ± 0.0000	20.8 ± 1.0
Raw-gate Relation†	–	✓	0.4243 ± 0.0000	0.4243 ± 0.0000	21.1 ± 0.8
Factorized-token†	✓	–	0.4164 ± 0.0080	0.4095 ± 0.0118	38.5 ± 2.6
MAGICCORE-LM	✓	✓	**0.3606 ± 0.0527**	**0.3532 ± 0.0514**	**50.6 ± 8.5**

LLM integration. Across five backbones (Yang et al., 2025a; Team et al., 2025; Abouelenin et al., 2025; Guo et al., 2025; Grattafiori et al., 2024) with three matched seeds each, MagicCore leads the sidecar–LLM comparison (Table 7). Macro OOD-BMAE falls 16.4% versus Raw and 9.2% versus Factorized, with positive paired CIs. The post-hoc frozen-sidecar ablation isolates an additional 0.83% macro OOD gain from the LLM path (Appendix B.8).

LLM neural-path cost. Neural-path speedup grows from $1.01\times$ ($L = 512$) to $6.18\times$ ($L = 4096$), with all backbones above parity by $L = 2048$ (Figures 3, 5). At $L = 4096$, inputs shrink from 4100 to 22 positions and mean peak memory falls 51.4%, with matched trainable counts. LLM Python front end (2.33 s MagicCore, 2.10 ms Raw) dominates full latency; Appendix Tables 9–12 give all 60 cells, separately from Compact E2E costs (Table 3).

Propagated-state confirmation. A separate three-seed replication reaches IID/OOD MAE 0.0666 ± 0.0179/0.1280 ± 0.0116 and OOD $R^2 = 0.630 \pm 0.058$ (full results and fit checks: Appendix B.3).

7 ANALYSIS: INTERFACE AND READER CONTRIBUTIONS

Removing absolute positions and adding relations each improves both split means at either setting of the other factor (Table 1); these are trends, not isolated significant effects. Reader ranking depends on schema: propagated-state DeepSets reaches MAE 0.1309, beating the relation-aware reader, whose intervals against vanilla, shuffled relations, and sidecar-only cross zero (Appendix B.4).

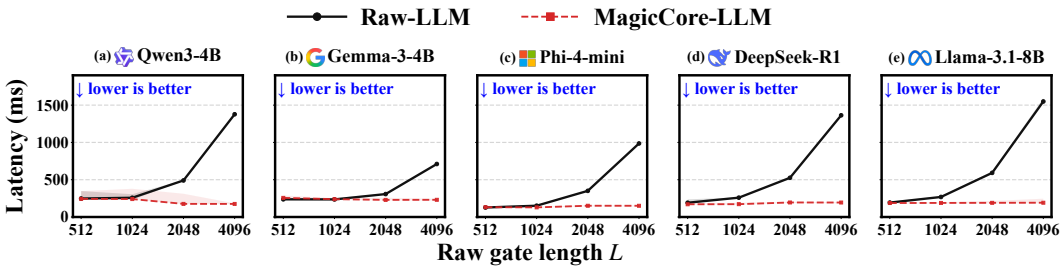


Figure 5: **Frozen neural-path latency across five local LLMs.** Matched batch, hardware, semantics, and readout. Black/red: Raw/MagicCore. Lines/bands: synchronized median/5th–95th wall-clock percentiles under neutral Clifford expansion. Front-end costs and protocol: Appendix A.7.

Table 7: **Fresh-blind transfer across five pretrained local-LLM backbones.** Three-seed mean ± SD; 45 candidates frozen before 2,000 IID + 2,000 OOD examples. Accuracy: Hard = five-class argmax; Sign = nonzero-target sign. **Best**; second best per backbone. Paired gains: Appendix A.

		Blind BMAE ↓		OOD accuracy (%) ↑	
Backbone	**Method**	**IID**	**Depth 512**	**Hard**	**Sign**
Qwen3-4B	Raw-LLM	0.4820±0.0001	0.4719±0.0024	23.0±0.4	49.9±0.6
	Factorized-LLM	0.4492±0.0018	0.4345±0.0017	41.0±0.6	56.1±0.5
	MAGICCORE-LM	**0.3890±0.0022**	**0.3944±0.0025**	**48.3±0.5**	**59.4±0.4**
Gemma-3-4B	Raw-LLM	0.4833±0.0016	0.4744±0.0046	23.6±0.5	50.6±0.4
	Factorized-LLM	0.4514±0.0021	0.4363±0.0023	40.5±0.6	55.3±1.2
	MAGICCORE-LM	**0.3908±0.0022**	**0.3957±0.0028**	**48.1±0.5**	**59.1±0.3**
Phi-4-mini	Raw-LLM	0.4824±0.0002	0.4731±0.0034	23.3±0.3	50.0±0.5
	Factorized-LLM	0.4483±0.0008	0.4338±0.0008	41.0±0.1	56.1±0.1
	MAGICCORE-LM	**0.3904±0.0005**	**0.3957±0.0006**	**48.0±0.1**	**58.8±0.5**
DeepSeek-R1-Qwen-7B	Raw-LLM	0.4829±0.0009	0.4732±0.0005	22.6±0.4	49.7±0.0
	Factorized-LLM	0.4491±0.0009	0.4342±0.0004	40.9±0.4	55.8±1.1
	MAGICCORE-LM	**0.3898±0.0027**	**0.3950±0.0033**	**48.5±0.4**	**59.2±0.4**
Llama-3.1-8B	Raw-LLM	0.4821±0.0003	0.4713±0.0020	23.7±0.5	50.4±0.3
	Factorized-LLM	0.4548±0.0016	0.4399±0.0009	41.1±1.0	55.9±1.4
	MAGICCORE-LM	**0.3910±0.0021**	**0.3966±0.0022**	**48.0±0.2**	**59.0±0.4**

8 LIMITATIONS

Compact OOD $R^2 = -0.084$ and the primary interval crossing zero constrain extrapolation and reader attribution. Continuous labels use observable mixing and algorithm transfer uses fixed angles. Compact generalization to larger support, general circuit angles, noise, and other initial states remains untested. Propagated fields permit analytic contraction. Appendix B.5 audits branch growth, truncation, and target concentration. Six additional frame-explicit seeds reuse held-out sets post hoc.

9 CONCLUSION

MAGICCORE-LM translates quantum algebra into unexecuted Compact instructions whose effects a reader learns. IID gains across readers (Table 1) and lower complete cost without retraining (Table 3) support this interface; frame-explicit, propagated-state, and LLM studies extend it. Stable input contracts let compilers improve without changing readers. More broadly, future scientific LLMs need more than raw inputs and model scale: domain expertise can identify structure to compute exactly, invariances and relations to expose, and responses to learn. This division of labor offers a practical design principle for adapting general-purpose models to specialized science. It also provides a reusable experimental pattern: hold the scientific input contract fixed, then evaluate improvements to learning and execution separately.

AI USE STATEMENT

Generative AI tools were used for research brainstorming, literature discovery, implementation and refactoring assistance, experiment orchestration, and polishing writing. The authors reviewed all AI-assisted work. Citations were checked against primary sources; generated code was covered by unit, algebraic, resume-integrity, and end-to-end tests; and all reported numbers were read from content-addressed experimental artifacts. The authors take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only synthetic quantum circuits and contains no human-subject, personal, or sensitive data.

REPRODUCIBILITY STATEMENT

Appendix A specifies the task, data generation, model dimensions, optimization, calibration, and bootstrap procedure.

REFERENCES

Scott Aaronson and Daniel Gottesman. Improved simulation of stabilizer circuits. *Physical Review A—Atomic, Molecular, and Optical Physics*, 70(5):052328, 2004.

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2(3):5, 2018.

Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, Shahnawaz Ahmed, Vishnu Ajith, M Sohaib Alam, Guillermo Alonso-Linaje, Bharath AkashNarayanan, Ali Asadi, et al. Pennylane: Automatic differentiation of hybrid quantum-classical computations. *arXiv preprint arXiv:1811.04968*, 2018.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Sergey Bravyi and David Gosset. Improved classical simulation of quantum circuits dominated by clifford gates. *Physical review letters*, 116(25):250501, 2016.

Sergey Bravyi, Graeme Smith, and John A Smolin. Trading classical and quantum computational resources. *Physical Review X*, 6(2):021043, 2016.

Sergey Bravyi, Dan Browne, Padraic Calpin, Earl Campbell, David Gosset, and Mark Howard. Simulation of quantum circuits by low-rank stabilizer decompositions. *Quantum*, 3:181, 2019.

Michael J Bremner, Richard Jozsa, and Dan J Shepherd. Classical simulation of commuting quantum computations implies collapse of the polynomial hierarchy. *Proceedings: Mathematical, Physical and Engineering Sciences*, pp. 459–472, 2011.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Simone Cantori and Sebastiano Pilati. Challenges and opportunities in the supervised learning of quantum circuit expectation values. *Physical Review E*, 112(1):015304, 2025.

Simone Cantori, David Vitali, and Sebastiano Pilati. Supervised learning of random quantum circuits via scalable neural networks. *Quantum Science and Technology*, 8(2):025022, 2023.

Juan Carrasquilla, Di Luo, Felipe Pérez, Ashley Milsted, Bryan K Clark, Maksims Volkovs, and Leandro Aolita. Probabilistic simulation of quantum circuits using a deep-learning architecture. *Physical Review A*, 104(3):032610, 2021.

Jeroen Dehaene and Bart De Moor. Clifford group, stabilizer states, and linear and quadratic operations over gf (2). *Physical Review A*, 68(4):042318, 2003.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Yuxuan Du, Min-Hsiu Hsieh, and Dacheng Tao. Efficient learning for linear properties of bounded-gate quantum circuits. *Nature Communications*, 16(1):3790, 2025.

Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm. *arXiv preprint arXiv:1411.4028*, 2014.

Daniel Gottesman. The heisenberg representation of quantum computers. *arXiv preprint quant-ph/9807006*, 1998.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Mark Howard and Earl Campbell. Application of a resource theory for magic states to fault-tolerant quantum computing. *Physical review letters*, 118(9):090501, 2017.

Mark Howard, Joel Wallman, Victor Veitch, and Joseph Emerson. Contextuality supplies the 'magic'for quantum computation. *Nature*, 510(7505):351–355, 2014.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, 2020.

Ali Javadi-Abhari, Matthew Treinish, Kevin Krsulich, Christopher J Wood, Jake Lishman, Julien Gacon, Simon Martiel, Paul D Nation, Lev S Bishop, Andrew W Cross, et al. Quantum computing with qiskit. *arXiv preprint arXiv:2405.08810*, 2024.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pp. 4582–4597, 2021.

Yuxiang Liu, Fanxu Meng, Lu Wang, Yi Hu, Zaichen Zhang, and Xutao Yu. Output prediction of quantum circuits based on graph neural networks. *arXiv preprint arXiv:2504.00464*, 2025.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Darya Martyniuk, Johannes Jung, Daniel Barta, and Adrian Paschke. Benchmarking quantum architecture search with surrogate assistance. *arXiv preprint arXiv:2506.06762*, 2025.

Johannes Jakob Meyer, Marian Mularski, Elies Gil-Fuster, Antonio Anna Mele, Francesco Arzani, Alissa Wilms, and Jens Eisert. Exploiting symmetry in variational quantum machine learning. *PRX quantum*, 4(1):010328, 2023.

Hakop Pashayan, Joel J Wallman, and Stephen D Bartlett. Estimating outcome probabilities of quantum circuits using quasiprobabilities. *Physical review letters*, 115(7):070501, 2015.

Qiskit Nature developers and contributors. Qiskit Nature, 2023.

Jonathan Romero, Ryan Babbush, Jarrod R McClean, Cornelius Hempel, Peter J Love, and Alán Aspuru-Guzik. Strategies for quantum computing molecular energies using the unitary coupled cluster ansatz. *Quantum Science and Technology*, 4(1):014008, 2019.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468, 2018.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

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.

Victor Veitch, SA Hamed Mousavian, Daniel Gottesman, and Joseph Emerson. The resource theory of stabilizer quantum computation. *New Journal of Physics*, 16(1):013009, 2014.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Hanrui Wang, Pengyu Liu, Jinglei Cheng, Zhiding Liang, Jiaqi Gu, Zirui Li, Yongshan Ding, Weiwen Jiang, Yiyu Shi, Xuehai Qian, et al. Quest: Graph transformer for quantum circuit reliability estimation. *arXiv preprint arXiv:2210.16724*, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Rui Yang, Ziruo Wang, Yuntian Gu, Tianyi Chen, Yitao Liang, and Tongyang Li. Qcircuitbench: A large-scale dataset for benchmarking quantum algorithm design, 2025b. URL https://arxiv.org/abs/2410.07961.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34:28877–28888, 2021.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

A REPRODUCIBILITY DETAILS

Compact learning and deployment are documented in Appendices B.1 and B.2. The following configuration and five-class protocol describe the frame-explicit extension (Table 2), not the Compact confirmation in Table 1.

A.1 FRAME-EXPLICIT CONFIGURATION

Table 8: **Frame-explicit configuration shared by the internal routes in Table 2.**

Component	**Setting**
Backbone	4 layers, width 200, 5 heads, FF multiplier 4
Internal parameters	1,964,005 per route
Optimizer	AdamW, batch 64, weight decay .01, clip 1
Scalar training	$2 \times 11{,}260$ successful updates
Five-class stage	5,630 updates; encoder/head LR $3e{-}5/3e{-}4$
Runtime	FP32 deterministic CUDA; dropout 0; seven matched seeds
Train/validation	36,000 / 4,000 examples
Final blind set	2,000 IID + $2{,}000 \times 3$ OOD records
Bootstrap	10,000 paired/hierarchical replicates

A.2 FRAME-EXPLICIT DATA AND LEAKAGE CONTROLS

Every IID and OOD parent pool has 400 parents per target stratum. Within each stratum, 200 use adjacent-inverse and 200 use compiled-identity rewrites. OOD parent renderings preserve query, target, and factorized bytes across depths. The final blind-set seed is 20260911. Generation rejects any collision in parent ID, semantic hash, orbit ID, or factorized-byte hash with the historical 58,000-record denylist or the other fresh pool. The builder persists shards and a manifest, reloads them, runs structural and numerical preflight, and exits before evaluation. Core circuits use $n \in [4, 10]$, $t \in [1, 8]$, and Pauli-query weight one to four. The evaluator verifies candidate, calibration-map, data, and decision-rule hashes before a single release and makes no model write after release.

A.3 PAULI ENCODING AND EXACT CHECKS

Frame-explicit records have 22 coordinates: ten padded x bits, ten padded z bits, a type (-1 for frame-X, 0 for frame-Z, $+1$ for a rotation), and a type-dependent sign/order feature. The last feature is a frame generator's Hermitian sign or $s_j\, j/t$ for rotation $j \in \{1, \ldots, t\}$. Sign and order thus share a type-dependent coordinate in this frame-explicit implementation. A separate 20-bit symplectic record encodes the query.

A Pauli $P = i^\kappa X^x Z^z$ is stored by binary vectors $x, z \in \mathbb{F}_2^n$ and a canonical sign. Commutation is the symplectic product $[P, Q] = x_P \cdot z_Q + z_P \cdot x_Q \pmod 2$. Clifford conjugation is implemented as tableau row updates (Dehaene & De Moor, 2003; Aaronson & Gottesman, 2004). Raw and factorized evaluators independently propagate the query in the Heisenberg picture; audited records additionally compare with complex128 state-vector evaluation. The largest audited formal-dataset error before model training was below 2.5×10^{-14}.

A.4 PUBLISHED-ARCHITECTURE ADAPTATIONS

The Cantori-style CNN uses a one-hot qubit-by-depth image with convolutional blocks, global pooling, and a Pauli-query column. The Liu-style GNN uses the circuit dependency DAG, 31-dimensional typed node features, four attention message-passing layers, mean pooling, and explicit query nodes. Widths are chosen before the blind set to yield 1.87M and 1.94M parameters, respectively, near the 1.96M internal budget. Both receive the same target head, optimizer-update count, train/validation examples, calibration family, and blind-set release as the internal routes; constant-zero is retained only as a sanity reference.

Comparison scope. The matched baselines share circuit/query inputs and a supervised expectation target. Classical shadows (Huang et al., 2020) instead consume measurements of the state, so a cost comparison would also require a shot budget. The POVM-based Transformer of Carrasquilla et al. (2021) represents a quantum state as a probability distribution and updates it through quasi-stochastic gate dynamics; its circuit-simulation optimization is distinct from evaluating a predictor frozen across unseen circuits. SQuASH (Martyniuk et al., 2025) benchmarks surrogate-assisted architecture search, while QCircuitNet (Yang et al., 2025b) evaluates quantum algorithm/code generation. These are related tasks, not interchangeable expectation-regression baselines; an adapted comparison would need its own matched targets, optimization, and compute budget.

A.5 Frame-explicit optimization and selection

The discrete frame-explicit task uses five centers $\mathcal{C} = \{-\sqrt{2}/2, -\sqrt{2}/4, 0, \sqrt{2}/4, \sqrt{2}/2\}$. For logits z, $p = \mathrm{softmax}(z)$ and $\mu = \sum_k p_k \mathcal{C}_k$, the objective is

$$\mathcal{L} = \mathrm{CE}(z, k_y) + \frac{(\mu - y)^2}{0.25}, \tag{6}$$

where k_y indexes the target center ($\mathcal{C}_{k_y} = y$) and 0.25 is the balanced target variance. All routes share this objective and calibrator; the continuous compact task instead uses Equation 4.

All core routes use batch size 64, deterministic FP32, gradient clipping at one, and AdamW with weight decay 0.01 (Loshchilov & Hutter, 2017). Scalar encoders receive two 11,260-update cosine cycles (Loshchilov & Hutter, 2016). The distributional stage then runs for 5,630 updates with encoder/head learning rates 3×10^{-5} and 3×10^{-4} and 282 warm-up steps, giving 28,150 successful updates per learned route. The scalar-risk calibrator for this discrete task is an odd, monotone piecewise-linear map with nine fixed magnitude knots on $[0, \sqrt{2}/2]$. Heights are fit by label-balanced L_1 risk. Five-fold parent-hash cross-fitting is used for model release, followed by a full-validation fit frozen for blind inference. These procedures and budgets are shared across methods. Class probabilities remain at temperature one; only the reported scalar is mapped.

The primary seed 3407 candidates were frozen before the two core held-out sets were generated. Seeds 4519, 8123, 4589, 5248, 5770, and 9098 were trained later under the same fixed configuration and evaluated on those already opened sets, providing post-hoc estimates of initialization variability. Hierarchical intervals resample both the seven training seeds and semantic parents, stratified by target label.

A.6 Local-LLM and complexity-OOD protocols

The local-backbone study covers Qwen3-4B, Gemma-3-4B, Phi-4-mini, DeepSeek-R1-Distill-Qwen-7B (abbreviated DeepSeek-R1-Qwen-7B), and Llama-3.1-8B. Table 7 uses three frozen 1,964,005-parameter sidecars, one per Raw, frame-explicit Factorized, and relation-aware MagicCore route. Each is trained on the 36,000-example core task for 28,150 updates; the same route's sidecar is then shared across all backbones and adaptation seeds. The adaptation set has 500 sign pairs (1,000 rendered records), and every run completes 500 updates at seeds 3407, 3408, and 3409. Sidecar states pass through LayerNorm(200), Linear$(200, 256)$, GELU, and Linear$(256, d_{\mathrm{LLM}})$ to become continuous prefixes. A learned REG token is processed by the full backbone's existing layers; g_θ is an affine five-logit head and $b = 0.25$ in Equation 5. The posterior uses the five centers above at temperature one, without an additional fitted scalar calibration.

QLoRA uses rank 8, scaling 16, zero adapter dropout, all linear target modules, and NF4 double quantization with bfloat16 compute; frozen sidecars remain FP32. One microbatch is one sign pair (two records), accumulated over four microbatches: eight records per update. AdamW uses learning rates 10^{-5} for LoRA/REG and 3×10^{-4} for the projection/head, zero weight decay, clipping at one, and 40 warmup steps followed by cosine decay to step 500. The adaptation loss is five-class CE plus posterior-mean MSE divided by 0.25 and a sidecar-posterior KL anchor, with weights $1, 1, 0.1$; all routes use the same objective. All 45 candidates are frozen before 2,000 IID and 2,000 depth-512 OOD records are generated; the release excludes 54,000 historical semantic parents and includes 84 complex128 state-vector audits. Macro OOD gains are 0.0773 over Raw (hierarchical 95% CI $[0.0595, 0.0949]$) and 0.0403 over Factorized ($[0.0207, 0.0598]$). Macro OOD-BMAE is 0.4728/0.4358/0.3955 for Raw/Factorized/MagicCore.

For the separate complexity stress test, sidecars train on 6,000 circuits with $t \in \{5, 7\}$; Qwen adaptation uses 500 selected records for eight exposures, with selection on 1,200 development examples. Clifford depth is 192, every additional $T/T^\dagger$ acts on the four active qubits, and all routes share an ordered six-center objective, eight data exposures, and three seeds. Nine frozen candidates are tested on 2,400 fresh examples balanced across $t \in \{9, 11, 13, 15\}$. MagicCore receives input-derived branch-state tokens and uses a frozen sidecar prediction plus a bounded Qwen logit residual. No target scalar is supplied, but the legacy coefficient/compatibility fields permit analytic contraction, as distinguished from the field-isolated audit in Appendix B.4.

Here the six centers are $\{-\sqrt{2}/4, -1/4, -\sqrt{2}/8, \sqrt{2}/8, 1/4, \sqrt{2}/4\}$, and the residual bound is $b = 0.5$, rather than Table 7's five-center bound of 0.25. The objective combines cross-entropy, variance-normalized posterior-expectation MSE, an ordinal cumulative-distribution penalty, and KL regularization toward the frozen sidecar posterior. Cross-entropy distinguishes six labels; the ordinal term compares cumulative probabilities at their five boundaries, and expectation MSE penalizes numerical misalignment. The four loss weights are $1, 1, 0.5, 0.1$, respectively; MSE is divided by $7/96$. All routes share these settings, the QLoRA ranks and optimizer schedule above, microbatch two with accumulation four, and 500 updates. No blind-set readout selection is performed.

Near-flat complexity-OOD baselines. The frozen predictions underlying Figure 4 vary little: within each seed and t stratum, prediction standard deviations range from 0.00029 to 0.00104 for Raw and 0.00080 to 0.00448 for Factorized, versus a target standard deviation of 0.27003. The six balanced centers give a zero-prediction BMAE of $(\sqrt{2}/4 + 1/4 + \sqrt{2}/8)/3 \approx 0.26011$. Both baselines lie near this reference, with small differences retained in Table 13; their similar bar heights reflect near-constant predictions, not clipped errors or a lower bound on achievable BMAE.

The continuous-observable stress test uses matched 2M-parameter Transformers trained for 600 updates on 960 independent cores at $t \in \{5, 7\}$. For anticommuting Paulis P, Q from the same core and a deterministic continuous angle α, the target is $y = \cos(\alpha)\langle P\rangle + \sin(\alpha)\langle Q\rangle$. Each record jointly encodes the circuit, $P, Q, \cos\alpha$, and $\sin\alpha$; one Transformer forward directly predicts y. Component expectations and the final contraction are not supplied as scalar inputs or component supervision; the legacy branch fields still make them analytically recoverable. All 960 train targets, 200 validation targets, and 600 fresh-blind targets are unique at 12 decimal places. The blind set contains 150 independent cores at each of $t \in \{9, 11, 13, 15\}$ and was generated only after all nine candidates passed the development gate and were frozen. Paired MAE gains are 0.0620 over Raw (95% CI $[0.0467, 0.0763]$) and 0.0769 over Factorized ($[0.0608, 0.0933]$).

For this single-forward stress test, the observable angle is $\alpha = 2\pi u$, where u is the unsigned big-endian integer in the first eight SHA-256 bytes divided by 2^{64}. The UTF-8 hash input joins angle seed, split name, semantic-core hash, left example ID, and right example ID with the literal separator `|`. Angle seeds are 20270921 (train), 20270922 (validation), and 20270923 (fresh blind). Anticommuting pairs are ordered by a separate `pair-rank` hash, and the first eligible pair per core is used; nonfinite, zero, duplicate-at-12-decimal, or legacy-center targets are excluded. This constructs observable mixtures, not continuously varying physical gate angles.

The original validation request was 120 balanced pairs per training complexity. After all nine candidates were byte-frozen—but before any validation metric or fresh-blind generation—the target-independent eligibility audit found 125 eligible cores at $t = 5$ and 111 at $t = 7$. We therefore recorded a protocol amendment to use the first 100 candidates at each t under the already fixed hash ordering. Checkpoints, models, seeds, selection ordering, gates, and blind generation remained unchanged; no candidate was retrained.

The semantic-family OOD study uses the same three matched 1.96M-parameter Transformer routes and seeds 3407, 4519, and 8123. It trains on 1,440 balanced examples drawn from random-brickwork, star-entangler, and ring-cluster generators, with $n \in [6, 10]$, five active qubits, $t \in \{5, 7\}$, and fixed Clifford depth 96. Selection uses 360 independent development records. Only after all nine candidates pass the predeclared development gate and are byte-frozen does the protocol generate 900 blind records balanced across the disjoint QAOA-phase-gadget, ladder-excitation, and echo-conjugation families. All routes have the same 900 successful updates, 20 data exposures, direct scalar head, and query features. Coefficients are computed from the input circuit/query rather than labels. As in the continuous test, the propagated payload permits analytic recovery, despite omitting component supervision and contracted scalar fields.

Table 9: **Token and CPU front-end audit.** Stream tokens precede complete neural positions; factorized entries show mean [min–max] over the five fixed semantic inputs. Raw includes route-specific tensorization; Factorized additionally includes exact factorization; MagicCore also materializes the relation tensor. Each cell contains 200 interleaved timed calls after warm-up. Within each L, **green** marks the lowest count/time and gray with underlining the next distinct value; equal token counts share the best rank.

		Neural input ↓	CPU front end (ms) ↓	
L	Interface	Stream / positions	p50	p95
512	Raw-LLM	512 / 516	**0.58**	**0.98**
	Factorized-LLM	**18 [14–22] / 22**	278.76	381.04
	MAGICCORE-LM	**18 [14–22] / 22**	289.61	386.14
1024	Raw-LLM	1024 / 1028	**0.81**	**1.24**
	Factorized-LLM	**18 [14–22] / 22**	577.26	778.24
	MAGICCORE-LM	**18 [14–22] / 22**	585.23	750.33
2048	Raw-LLM	2048 / 2052	**1.23**	**1.70**
	Factorized-LLM	**18 [14–22] / 22**	1169.13	1523.90
	MAGICCORE-LM	**18 [14–22] / 22**	1184.11	1513.15
4096	Raw-LLM	4096 / 4100	**2.10**	**2.77**
	Factorized-LLM	**18 [14–22] / 22**	2338.88	3024.97
	MAGICCORE-LM	**18 [14–22] / 22**	2334.41	2980.51

The third-party standard-algorithm audit reuses those nine byte-frozen candidates without retraining, selection, or adaptation. Qiskit 2.5.2 constructs MaxCut QAOA at $p = 1$ and commuting IQP circuits; Qiskit Nature 0.7.2 constructs a minimal UCCSD ansatz and maps its occupied-to-virtual excitations with `JordanWignerMapper`. A separate exact compiler maps the resulting operations to $H/S/S^\dagger/\mathrm{CX}/T/T^\dagger$ without approximate synthesis. PennyLane 0.43.3 `default.qubit` independently labels every component, with maximum disagreement 1.45×10^{-14} against the exact Clifford+T evaluator. The generated pool contains 660 independent cores; a source-hashed balanced rule selects 600 records (200 per family), with $n \in [6, 10]$ and four active qubits. QAOA, IQP, and UCCSD contain respectively 8, 16, and 12 non-Clifford rotations. Every target is unique at 12 decimals, none matches a legacy center, and semantic hashes are disjoint from source training, validation, and prior blind data. An evaluation-only compatibility amendment preceded every model forward and preserved all candidate and blind-data hashes; its full receipt is retained with the result artifact. This audit measures transfer under the combined shift in software generator, circuit family, active-qubit count, and non-Clifford complexity.

A.7 TRAINING-FREE EFFICIENCY AUDIT

The latency extension, prospectively specified and hash-frozen before benchmarking, adds $L = 2048, 4096$ to the initial $512, 1024$ sweep. Geometric-mean neural-path speedup is $1.01\times$, $1.22\times$, $2.36\times$, and $6.18\times$ at $L = 512, 1024, 2048, 4096$, respectively, averaging the five paired Raw/MagicCore latency ratios.

We reuse the five semantic inputs, 15 seed-3407 checkpoints, and all 60 frozen neural timing cells from the sealed long-context benchmark; no model is trained and no validation or blind label is opened. CPU front ends are timed with one thread using 200 interleaved calls per method and length. Raw timing includes route-specific tensorization. Factorized timing recomputes the exact Clifford-frame/magic-rotation factorization before tensorization, and MagicCore additionally materializes the relation tensor. The sidecar contributes three common pool/query/metadata states and the LLM one readout state, so complete neural positions equal stream length plus four. Checkpoint tensors are counted directly and authenticated against the candidate-freeze manifest.

Table 10: **Frozen neural-path latency for all 60 cells.** Entries are synchronized batch-one wall-clock medians in milliseconds. The p05/p95 samples and CUDA-event timings are retained in the content-addressed CSV/JSON artifact. **Green** marks the lowest value and gray with underlining the next distinct value within each backbone and column; ties at the displayed precision share a rank.

		Neural-path latency (ms) ↓			
Backbone	**Interface**	$L = 512$	$L = 1024$	$L = 2048$	$L = 4096$
Qwen3-4B	Raw-LLM	247.6	258.0	490.2	1377.5
	Factorized-LLM	**240.2**	**236.0**	**173.8**	191.2
	MAGICCORE-LM	243.2	240.5	175.9	**175.3**
Gemma-3-4B	Raw-LLM	**237.0**	**235.8**	307.4	710.7
	Factorized-LLM	247.9	247.5	235.1	**230.1**
	MAGICCORE-LM	256.5	238.9	**229.3**	230.7
Phi-4-mini	Raw-LLM	**127.2**	151.5	350.8	986.0
	Factorized-LLM	131.3	131.5	151.5	**148.5**
	MAGICCORE-LM	131.5	**128.9**	**150.9**	150.5
DeepSeek-R1-Qwen-7B	Raw-LLM	191.3	259.1	525.8	1363.6
	Factorized-LLM	**166.6**	**167.8**	**192.0**	**192.5**
	MAGICCORE-LM	170.5	172.1	194.7	194.1
Llama-3.1-8B	Raw-LLM	193.2	268.6	591.9	1549.1
	Factorized-LLM	207.1	193.2	212.8	214.9
	MAGICCORE-LM	**187.3**	**188.7**	**190.1**	**191.5**

Table 11: **Peak inference memory for all 60 cells.** Entries are peak allocated CUDA memory in MiB; they include the quantized backbone and the complete frozen interface. **Green** marks the lowest value and gray with underlining the next distinct value within each backbone and column; ties at the displayed precision share a rank.

		Peak allocated CUDA memory (MiB) ↓			
Backbone	**Interface**	$L = 512$	$L = 1024$	$L = 2048$	$L = 4096$
Qwen3-4B	Raw-LLM	3694	4085	5305	9304
	Factorized-LLM	**3444**	**3444**	**3443**	**3443**
	MAGICCORE-LM	**3444**	**3444**	**3443**	**3443**
Gemma-3-4B	Raw-LLM	4771	5066	5720	7270
	Factorized-LLM	**4550**	**4550**	**4550**	**4550**
	MAGICCORE-LM	4551	4551	**4550**	**4550**
Phi-4-mini	Raw-LLM	4363	4726	5738	9033
	Factorized-LLM	**4119**	**4119**	**4119**	**4119**
	MAGICCORE-LM	**4119**	**4119**	**4119**	**4119**
DeepSeek-R1-Qwen-7B	Raw-LLM	5851	6264	7345	10941
	Factorized-LLM	**5560**	**5560**	**5560**	**5560**
	MAGICCORE-LM	**5560**	**5560**	**5560**	**5560**
Llama-3.1-8B	Raw-LLM	6034	6479	7809	12135
	Factorized-LLM	**5689**	**5689**	**5689**	**5689**
	MAGICCORE-LM	5690	5690	**5689**	**5689**

A.8 COMPLEXITY-OOD NUMERICAL RESULTS

At the endpoints, the paired Raw-relative BMAE gains are 0.2464 at $t = 9$ (95% CI $[0.2427, 0.2502]$) and 0.1875 at $t = 15$ ($[0.1661, 0.2156]$).

Table 12: **Exact parameter audit for all frozen interfaces.** Sidecar is the stage-one matched 1.964M-parameter encoder. Stage 2 is LoRA + projection + readout + regression token. Across stages is their sum; the stages are optimized separately and all parameters are frozen during this audit. Counts are identical across interfaces within every backbone. Totals use integer counts; displayed components are rounded independently.

		Component parameters (M)					Totals (M)	
Backbone	**Interface**	**Sidecar**	**LoRA**	**Proj.**	**Readout**	**Reg. token**	**Stage 2**	**Across stages**
Qwen3-4B	Raw-LLM	1.964	16.515	0.710	0.013	0.003	17.240	19.204
	Factorized-LLM	1.964	16.515	0.710	0.013	0.003	17.240	19.204
	MAGICCORE-LM	1.964	16.515	0.710	0.013	0.003	17.240	19.204
Gemma-3-4B	Raw-LLM	1.964	19.249	0.710	0.013	0.003	19.974	21.938
	Factorized-LLM	1.964	19.249	0.710	0.013	0.003	19.974	21.938
	MAGICCORE-LM	1.964	19.249	0.710	0.013	0.003	19.974	21.938
Phi-4-mini	Raw-LLM	1.964	11.534	0.841	0.015	0.003	12.394	14.358
	Factorized-LLM	1.964	11.534	0.841	0.015	0.003	12.394	14.358
	MAGICCORE-LM	1.964	11.534	0.841	0.015	0.003	12.394	14.358
DeepSeek-R1-Qwen-7B	Raw-LLM	1.964	20.185	0.973	0.018	0.004	21.180	23.144
	Factorized-LLM	1.964	20.185	0.973	0.018	0.004	21.180	23.144
	MAGICCORE-LM	1.964	20.185	0.973	0.018	0.004	21.180	23.144
Llama-3.1-8B	Raw-LLM	1.964	20.972	1.105	0.020	0.004	22.101	24.065
	Factorized-LLM	1.964	20.972	1.105	0.020	0.004	22.101	24.065
	MAGICCORE-LM	1.964	20.972	1.105	0.020	0.004	22.101	24.065

Table 13: **Qwen3-4B extrapolation to greater non-Clifford complexity.** All routes use the same Qwen backbone, QLoRA budget, objective, and sidecar/residual readout; MagicCore additionally supplies propagated branch states. Candidates are trained and selected only at $t \in \{5, 7\}$, then hash-frozen before generating a fresh balanced blind set at $t \in \{9, 11, 13, 15\}$. Entries are mean $\pm$ sample standard deviation over three matched training seeds; "Avg." pools the four blind complexities within each seed. The paired aggregate BMAE gain of MAGICCORE-LM over Raw-LLM is 0.2194 (95% CI $[0.2097, 0.2332]$), and over Factorized-LLM is 0.2194 ($[0.2095, 0.2333]$). **Green** marks the best mean and gray with underlining the next distinct mean in each column; ties at the displayed precision share a rank.

	Balanced MAE ↓					Exact-6 ↑
Interface	$t = 9$	$t = 11$	$t = 13$	$t = 15$	**Avg.**	**Accuracy (%)**
Raw-LLM	.26011 ± .00001	.26012 ± .00001	.26011 ± .00001	.26010 ± .00002	.26011 ± .00001	16.43 ± .28
Factorized-LLM	.26013 ± .00005	.26007 ± .00008	.26010 ± .00003	.26014 ± .00007	.26011 ± .00000	16.92 ± .22
MAGICCORE-LM	**.01368 ± .00346**	**.03012 ± .00688**	**.04665 ± .01439**	**.07259 ± .02608**	**.04076 ± .01255**	**65.85 ± 5.70**

A.9 METRICS AND RELEASE RULE

Core BMAE is the equal mean of absolute error over the five label strata; OOD-BMAE additionally averages the three counterfactual depths. With $\mathcal{B}$ denoting all learned comparators and constant-zero, the frozen statistic is

$$\Delta^* = \min_{b \in \mathcal{B}} [\text{OOD-BMAE}(b) - \text{OOD-BMAE}(\text{MAGICCORE-LM})].$$

Each of 10,000 bootstrap replicates resamples the 2,000 OOD semantic parents, retains their three depths as one unit, and stratifies by label. The declared decision is GO iff $\Delta^* \geq 0.020$ and the paired

95% interval has positive lower endpoint. IID error, exact class accuracy, per-label cells, and cost are secondary diagnostics.

B Additional results

B.1 Compact learning-interface confirmation

Instruction inputs and shared readout. Unlike the propagated-state studies, Table 1 exposes ordered rotation instructions and two Pauli queries. For $U = MC$, exact conjugation gives $P'_j = C^\dagger P_j C$ and $Q' = C^\dagger Q C$, so prediction can use $M' = C^\dagger M C$ on $|0^n\rangle$. Inactive qubits are removed and the remaining columns are canonically ordered by their complete rotation/query incidence signatures. Specifically, a qubit's signature concatenates its (x, z) bits over the ordered rotations, then query 1 and query 2. All-zero columns are dropped; active signatures are sorted in descending lexicographic order. Equal signatures retain ascending original qubit index (a stable sort); swapping identical columns leaves the encoded records unchanged. Rotation order itself is never sorted. The resulting $t + 2$ records separate Pauli bits, rotation signs, physical order, and query identities. No non-Clifford branch expansion or terminal contribution is supplied. Relations encode symplectic parity, real and imaginary Pauli-product phase, and signed rotation order. Absolute serialization positions and relations form a complete 2×2 Transformer control; physical order remains a field in every compact arm. Compiler-normalized Raw uses exact Qiskit Clifford-block resynthesis, retains $T/T^\dagger$ boundaries, and applies a gate-count guard and disjoint inverse cancellation.

For positive Hermitian Paulis P_i, write $P_i P_j = \omega_{ij} P_{i\oplus j}$, with $\omega_{ij} \in \{1, -1, i, -i\}$ and $i \oplus j$ denoting XOR of the binary Pauli masks. The compact relation channels are $[P_i, P_j]$, $\mathrm{Im}\,\omega_{ij}$, signed order difference, and $\mathrm{Re}\,\omega_{ij}$; they are computed locally from records. The frame-explicit core variant instead uses parity, support Jaccard overlap, signed magic order, and equal type. Deterministic input/output checks cover compiler equivalence, same-input target consistency, permutation/padding behavior, and shared readout linearity. These checks validate the contract, not out-of-distribution predictive accuracy.

For Equation 3, the four channels are routed to five heads by the first four columns of the orthonormal DCT-II basis. With zero-based head $h = 0, \ldots, 4$ and channel $r = 0, \ldots, 3$ in the listed order,

$$D_{hr} = a_r \cos\left(\frac{\pi(h + 1/2)r}{5}\right), \qquad a_0 = 1/\sqrt{5}, \quad a_{r>0} = \sqrt{2/5}, \qquad D^\top D = I_4.$$

It is constructed in float64, cast to the attention dtype/device, and held fixed with unit relation scale. Channelwise $\tanh$ precedes this projection; neither step adds parameters.

All readers make one forward pass. A shared head estimates two query responses, projected smoothly into the unit disk as $r = u/\sqrt{1 + \|u\|_2^2}$. The final scalar is $\hat{y} = w_1 r_1 + w_2 r_2$, with $w = (\cos\alpha, \sin\alpha)$ adjusted for canonical query signs; known observable weights enter this final readout, not the response backbone. Training uses only joint-target MSE, with deterministic TRAIN-only angle augmentation shared across methods. The six Transformer arms and Set Transformer have 1,963,201 trainable parameters; DeepSets has 1,973,297 (0.51% more). For finite u, the shared head lies strictly inside the disk and approaches its boundary as $\|u\|$ grows. Weight decay regularizes these parameters but imposes no hard norm ceiling.

Fixed development choices and fresh confirmation. One development seed evaluates four equal-budget optimizer profiles: constant/cosine schedules crossed with weight decay 0.01/0.1, all at learning rate 3×10^{-4}. Its original multi-condition gate is not met. A subsequent recorded amendment fixes the selected profile for each method and authorizes three new seeds, 196613, 216091, and 262147, independently of new performance. Raw and absolute-position compact vanilla use cosine/0.1; normalized Raw uses cosine/0.01; position-free compact vanilla uses constant/0.1; the remaining four readers use constant/0.01. Neither the original failed result nor its thresholds are changed.

All 24 trajectories use 3,840 independent TRAIN cores and 200 validation cores at $t = 5, 7$, with $n \in [4, 10]$ and four active qubits, batch 32, 7,200 updates (60 exposures), and matched within-seed example/augmentation schedules. Checkpoints at exposures 5, 10, 20, 30, 40, and 60 are evaluated on validation; minimum MAE selects the checkpoint, with ties favoring the earlier step. All selected

Table 14: **Compact confirmation: paired MAE comparisons.** Positive differences favor MagicCore. Intervals resample paired seeds and independent cores within t strata (5,000 draws). The starred OOD comparison is primary; all others are secondary with unadjusted intervals.

	Fresh IID		Fresh OOD	
Comparator	Δ MAE	95% CI	Δ MAE	95% CI
Raw	+0.02977	[+0.01321, +0.05001]	+0.01052	[-0.00131, +0.02411]
Compiler-normalized Raw	+0.03507	[+0.01726, +0.05535]	+0.01591	[+0.00332, +0.03101]
Compact: P=1, R=0	+0.01904	[+0.00614, +0.03379]	+0.01616	[+0.00202, +0.03396]
Compact: P=0, R=0	+0.01034	[-0.00396, +0.02489]	+0.00254*	[-0.00594, +0.01224]
Compact: P=1, R=1	+0.00510	[-0.00374, +0.01379]	+0.00702	[-0.00244, +0.01689]
Compact DeepSets	+0.01397	[+0.00353, +0.02444]	+0.00432	[-0.00341, +0.01246]
Compact Set Transformer	+0.01253	[+0.00216, +0.02332]	+0.00496	[-0.00279, +0.01278]
Zero	+0.02301	[+0.01103, +0.03521]	+0.00413	[-0.00462, +0.01315]
TRAIN mean	+0.02304	[+0.01100, +0.03527]	+0.00409	[-0.00468, +0.01311]

Table 15: **All three compact-confirmation seeds.** Validation-selected MAE; each column uses the same blind examples. **Best** and second best among all eight methods.

	Fresh IID			Fresh OOD		
Input / reader	196613	216091	262147	196613	216091	262147
Raw	0.2284	0.2477	0.2270	0.2336	0.2483	0.2313
Compiler-normalized Raw	0.2316	0.2325	0.2550	0.2375	0.2379	0.2540
Compact: P=1, R=0	0.2200	0.2200	0.2309	0.2371	0.2338	0.2593
Compact: P=0, R=0	**0.2056**	0.2206	0.2186	**0.2264**	0.2278	0.2350
Compact: P=1, R=1	0.2122	0.2090	0.2079	0.2283	0.2370	0.2374
MagicCore	0.2091	**0.2017**	**0.2031**	0.2294	**0.2265**	**0.2258**
Compact DeepSets	0.2195	0.2187	0.2176	0.2307	0.2327	0.2312
Compact Set Transformer	0.2171	0.2171	0.2172	0.2344	0.2302	0.2320

and fixed-exposure weights are frozen before generating 400 IID cores and 800 OOD cores, the latter balanced over $t = 9, 11, 13, 15$. Prior proposed cores are excluded. Development and confirmation seeds are never pooled, and blind outcomes are not used to select an exposure.

Outcomes and uncertainty. The primary comparison is fresh-OOD MAE of position-free compact vanilla minus MagicCore: 0.00254, 95% paired interval $[-0.00594, 0.01224]$. Other comparisons and IID results are secondary, with unadjusted intervals. MagicCore leads average MAE on both splits; its IID/OOD R^2 is $0.092/-0.084$ and prediction/target standard-deviation ratio is 0.568/0.593. Zero-reference MAE is 0.2276/0.2314. Table 14 retains every comparator, and Table 15 shows all three seeds. The result archive additionally contains all six exposure views and every per-t metric.

Direct reader–baseline comparisons and seed sensitivity. We additionally analyze the existing selected predictions in eight direct comparisons: position-free compact vanilla, DeepSets, Set Transformer, and the position-matched P=1,R=0 Compact control, each versus Raw and compiler-normalized Raw, on both splits. This is a post-hoc secondary analysis, with no new training, inference, blind data, or checkpoint selection. The original primary endpoint and its outcome remain unchanged. For each paired seed/core, the difference is baseline absolute error minus reader absolute error. We directly resample these differences, rather than subtracting existing interval endpoints. The original crossed bootstrap draws three paired training seeds and an independent core resample within each t stratum; the same core indices apply to every sampled seed and both methods. Stratum counts stay fixed (5,000 draws, RNG seed 932118). All 16 intervals in Table 16 are unadjusted, pointwise 95% intervals, not a simultaneous family-wise claim.

Table 16: **Direct Compact–Raw comparisons.** Post-hoc secondary analysis of the frozen selected predictions. $\Delta = \mathrm{MAE}_{\mathrm{baseline}} - \mathrm{MAE}_{\mathrm{reader}}$; positive favors Compact. All 16 intervals are pointwise, unadjusted 95% crossed seed/core bootstrap intervals (5,000 draws). **Green** marks intervals strictly above zero, not a ranking or family-wise test. P/R denote absolute positions/algebraic relations; $^{\mathrm{T}}$ denotes Transformer.

		Fresh IID		Fresh OOD	
Compact reader	**Baseline**	Δ MAE	95% CI	Δ MAE	95% CI
Compact$^{\mathrm{T}}$ (P=0,R=0)	Raw	+0.01943	[+0.00621, +0.03307]	+0.00798	[-0.00437, +0.02099]
	Norm. Raw	+0.02473	[+0.01080, +0.04077]	+0.01337	[+0.00639, +0.02235]
Compact DeepSets	Raw	+0.01581	[+0.00430, +0.03221]	+0.00621	[-0.00191, +0.01667]
	Norm. Raw	+0.02111	[+0.00695, +0.03939]	+0.01159	[+0.00226, +0.02361]
Compact Set Transformer	Raw	+0.01724	[+0.00514, +0.03342]	+0.00557	[-0.00452, +0.01834]
	Norm. Raw	+0.02254	[+0.00896, +0.03985]	+0.01095	[+0.00103, +0.02300]
Additional position-matched Transformer control					
Compact$^{\mathrm{T}}$ (P=1,R=0)	Raw	+0.01073	[-0.00724, +0.02991]	-0.00564	[-0.02697, +0.01462]
	Norm. Raw	+0.01603	[+0.00393, +0.03139]	-0.00026	[-0.01072, +0.00811]

The three position-free readers have positive IID intervals against both baselines and positive OOD intervals against normalized Raw; their OOD intervals against Raw cross zero. The position-matched control has a positive interval only for IID versus normalized Raw. Table 17 reports every seed difference, their sample SD, win count, and the range of leave-one-seed-out means. This range is descriptive, not a confidence interval; the many test cores do not replace independent training runs. Set Transformer's OOD Raw-relative mean gain, for example, accompanies only one improving seed out of three. The position-free comparison changes positional treatment and the set-reader comparisons also change architecture, supporting usefulness across readers rather than exclusive attribution to representation alone. All per-method optimizer choices remain frozen. The analysis archive reproduces the original MAE/R^2 and 18 existing paired intervals before reporting these direct comparisons.

B.2 Frozen Compact compiler and deployment audit

This retrospective audit reuses the validation-selected checkpoints and already released 400 IID / 800 OOD cores of Table 1, without training, reselection, or new blind generation. Raw and compiler-normalized Raw retain their own three frozen readers. Four Compact implementations share the same three MagicCore readers: reference CPU, packed-bit CPU with the original encoder, packed CPU with GPU encoding, and fused GPU compilation/encoding. Every reader has 1,963,201 parameters. The full comparison is in Table 18; Table 19 gives paired cost intervals.

Exact implementation change. The packed compiler maintains the inverse Clifford-prefix images of X_q, Z_q as bit masks and phases modulo four. Each $T/T^\dagger$ records the current inverse image of Z_q; the final inverse Clifford transforms the two query Paulis. The same inactive-column removal and canonical ordering then produce the frozen $t + 2$ records. The CUDA version executes one fused sequential compiler per circuit, with circuits parallelized across threads, followed by a fused encoder. Neither implementation propagates non-Clifford branches or contracts an expectation. All 1,200 compiled keys and input tensors are exactly equal to the reference implementation; all 21,600 full-data predictions are retained. Same-batch prediction differences are zero. Under the unchanged TF32 settings, regrouping batches produces a maximum difference of 3.16×10^{-4}, recorded separately from compiler equivalence. Full-data errors keep the original batch-32 grouping. On OOD inputs, packed CPU reduces median compilation from 113.16ms to 0.249ms relative to the reference CPU implementation.

Matched measurement boundaries. A fixed, target-independent hash order selects 32 IID cores (16 per t) and 32 OOD cores (eight per t), shared by every path and seed at batch sizes 1, 8, and 32. Measurements use one RTX 4090 at a recorded 330W limit, two CPU threads, FP32 with

Table 17: **Seed-level sensitivity of the direct comparisons.** Each seed column is baseline-minus-reader MAE on the same blind cores. SD is the sample standard deviation of the three paired differences; wins count positive differences. LOO is the range of means after omitting each seed in turn: a descriptive sensitivity range, *not* a confidence interval. Three training seeds remain three independent training runs.

Compact reader	Baseline	196613	216091	262147	SD	Wins	LOO range
Fresh IID							
Compact^{T} (P=0,R=0)	Raw	+0.02275	+0.02715	+0.00839	0.00981	3/3	[+0.0156, +0.0250]
	Norm. Raw	+0.02596	+0.01189	+0.03634	0.01227	3/3	[+0.0189, +0.0312]
Compact DeepSets	Raw	+0.00888	+0.02907	+0.00947	0.01149	3/3	[+0.0092, +0.0193]
	Norm. Raw	+0.01208	+0.01380	+0.03743	0.01416	3/3	[+0.0129, +0.0256]
Compact Set Transformer	Raw	+0.01128	+0.03062	+0.00981	0.01162	3/3	[+0.0105, +0.0210]
	Norm. Raw	+0.01449	+0.01536	+0.03776	0.01319	3/3	[+0.0149, +0.0266]
Compact^{T} (P=1,R=0)	Raw	+0.00836	+0.02768	-0.00385	0.01590	2/3	[+0.0023, +0.0180]
	Norm. Raw	+0.01156	+0.01242	+0.02411	0.00701	3/3	[+0.0120, +0.0183]
Fresh OOD							
Compact^{T} (P=0,R=0)	Raw	+0.00716	+0.02050	-0.00371	0.01212	2/3	[+0.0017, +0.0138]
	Norm. Raw	+0.01110	+0.01003	+0.01897	0.00488	3/3	[+0.0106, +0.0150]
Compact DeepSets	Raw	+0.00287	+0.01565	+0.00010	0.00829	3/3	[+0.0015, +0.0093]
	Norm. Raw	+0.00681	+0.00518	+0.02278	0.00972	3/3	[+0.0060, +0.0148]
Compact Set Transformer	Raw	-0.00080	+0.01816	-0.00066	0.01091	1/3	[-0.0007, +0.0087]
	Norm. Raw	+0.00314	+0.00770	+0.02201	0.00985	3/3	[+0.0054, +0.0149]
Compact^{T} (P=1,R=0)	Raw	-0.00352	+0.01451	-0.02791	0.02129	1/3	[-0.0157, +0.0055]
	Norm. Raw	+0.00042	+0.00405	-0.00524	0.00468	2/3	[-0.0024, +0.0022]

the frozen TF32 settings, two warmups, and five repetitions. Compiled-result and encoder caches are cleared outside every timed call; weights, CUDA runtime, and the compiled kernel remain warm. Kernel initialization is recorded separately. Independently timed E2E spans in-memory raw-gate requests through compilation, encoding, transfers, neural inference, and CPU predictions. Separate instrumented passes measure the stages; their medians need not sum to the E2E median. GPU Compile includes raw-gate packing and host-to-device transfer; CPU+GPU Encode includes compact-state packing and transfer. All residual transfer and neural costs enter E2E.

Records are structured gate/rotation/query tokens, not text-tokenizer subwords; three internal prefix tokens are additional. Reported GPU peaks summarize per-batch PyTorch-allocated peaks, including resident weights and temporary tensors but excluding driver/JIT-code allocations. Reserved-memory and maximum-request peaks, batch-eight outcomes, neural stage timings, and timing-cohort errors are retained in the machine-readable archive. Full-data MAE and timing-cohort MAE are distinct: the latter is IID/OOD 0.2352/0.2492 for Raw, 0.2595/0.2590 for normalized Raw, and 0.2122/0.2104 for all Compact paths.

Intervals use 3,000 bootstrap draws of paired seeds and timing batches after collapsing repetitions. At batch 32, each split has one timing batch, so its interval reflects variation across the three seed runs rather than independently repeated hardware sessions. GPU is not uniformly fastest: the hybrid CPU+GPU encoder has the lowest IID single-query latency, 5.39ms. The four-row main table is a representative deployment comparison, with ranks among its displayed rows; the six-row appendix retains both omitted implementations. These costs compare the measured frozen deployments, including the existing Raw encoders, on the audited distribution. The current CUDA kernel supports up to ten qubits and 32 rotations.

Table 18: **Complete six-path Compact compiler comparison.** The same three frozen MagicCore readers serve all Compact paths; Raw controls retain their own checkpoints. MAE: full-data mean ± sample SD. Times are medians; stage columns use $B = 1$, and $B = 32$ is whole-batch E2E latency. **Lowest** and next distinct displayed errors/costs within each split. Stage columns are unranked; their transfer boundaries differ.

			Front-end stages (ms)		E2E latency (ms) ↓		GPU peak
Input / compiler	Records	MAE ↓	Compile	Encode	$B = 1$	$B = 32$	MiB ↓ $B = 32$
(a) IID $t \in \{5, 7\}$							
Raw	200.0	0.2344 ±0.0116	0.032	2.515	8.28	126.59	141.58
Normalized Raw	161.2	0.2397 ±0.0133	4.083	2.057	11.61	261.27	127.16
Compact / reference CPU	8.0	**0.2046 ±0.0039**	98.384	0.889	104.44	3076.36	**44.10**
Compact / packed CPU	8.0	**0.2046 ±0.0039**	0.212	0.703	6.33	40.18	**44.10**
Compact / CPU + GPU enc.	8.0	**0.2046 ±0.0039**	0.209	0.308	**5.39**	16.08	**44.10**
Compact / fused GPU	8.0	**0.2046 ±0.0039**	0.675	0.099	6.47	**15.78**	**44.10**
(b) Complexity OOD $t \in \{9, 11, 13, 15\}$							
Raw	206.0	0.2377 ±0.0092	0.040	3.691	9.33	135.06	146.41
Normalized Raw	170.4	0.2431 ±0.0094	5.269	2.163	12.88	302.49	132.32
Compact / reference CPU	14.0	**0.2272 ±0.0019**	113.159	1.108	120.74	3504.75	**47.20**
Compact / packed CPU	14.0	**0.2272 ±0.0019**	0.249	0.979	6.98	48.40	**47.20**
Compact / CPU + GPU enc.	14.0	**0.2272 ±0.0019**	0.238	0.307	5.68	16.78	**47.20**
Compact / fused GPU	14.0	**0.2272 ±0.0019**	0.673	0.091	**5.18**	**16.04**	**47.20**

Table 19: **Paired end-to-end speedup relative to Raw.** Geometric means and 95% intervals, not ratios of the displayed latency medians. Values above one favor the listed path. Repetitions are collapsed before resampling paired seeds and timing batches (3,000 draws).

	IID		OOD	
Input / compiler	$B = 1$	$B = 32$	$B = 1$	$B = 32$
Raw	1.00× [1.00, 1.00]	1.00× [1.00, 1.00]	1.00× [1.00, 1.00]	1.00× [1.00, 1.00]
Normalized Raw	0.74× [0.68, 0.80]	0.52× [0.47, 0.62]	0.74× [0.70, 0.79]	0.43× [0.34, 0.51]
Compact / reference CPU	0.09× [0.07, 0.10]	0.04× [0.03, 0.04]	0.08× [0.07, 0.09]	0.04× [0.04, 0.04]
Compact / packed CPU	1.27× [1.08, 1.47]	3.39× [3.03, 4.12]	1.32× [1.20, 1.45]	2.83× [2.42, 3.29]
Compact / CPU + GPU enc.	1.50× [1.35, 1.68]	8.55× [7.56, 10.89]	1.60× [1.39, 1.82]	7.77× [7.42, 8.16]
Compact / fused GPU	1.29× [1.08, 1.55]	7.94× [7.16, 8.72]	1.68× [1.48, 1.88]	9.15× [7.76, 11.60]

B.3 Propagated-state continuous-task confirmation

Development choices fixed before confirmation. A bounded development search at seed 3407 evaluates the same three learning rates ($10^{-4}, 3\times10^{-4}, 10^{-3}$) and exposure checkpoints for each

Table 20: **Prospective confirmation after bounded development tuning.** Matched 1.96M-parameter Transformers, three new seeds, 60 exposures. Development-selected checkpoints are frozen before 400 IID and 800 OOD cores are generated (Appendix B.3). Mean ± sample SD; **best**; second best. Zero-reference MAE: 0.2428/0.2329 (IID/OOD).

	Fresh IID ($t = 5, 7$)		**Fresh OOD** ($t = 9, 11, 13, 15$)	
Interface	MAE ↓	R^2 ↑	MAE ↓	R^2 ↑
Raw	0.2432 ± 0.0021	-0.036 ± 0.035	0.2379 ± 0.0031	-0.077 ± 0.052
Factorized	0.2616 ± 0.0153	-0.305 ± 0.235	0.2533 ± 0.0177	-0.312 ± 0.259
MAGICCORE-LM	**0.0666 ± 0.0179**	**0.891 ± 0.055**	**0.1280 ± 0.0116**	**0.630 ± 0.058**

interface. Its selected learning rates, $10^{-4}/10^{-4}/3\times10^{-4}$ for Raw/Factorized/MagicCore, are fixed before training the new seeds 17389, 29411, and 63127. The unchanged four-layer, five-head Transformer has width 200 and 1,963,201 parameters; the scalar tanh/MSE head, 960 training cores, and 200 development-validation cores at $t \in \{5, 7\}$ are shared. MagicCore receives propagated Pauli bits, coefficients, and joint-query tags, with explicit compatibility/contribution columns omitted. The state schema retains the algebraic reconstruction described in Appendix B.4.

Matched training and prospective release. Each seed uses identical initial weights and example schedules across interfaces. All nine trajectories complete 1,800 FP32 AdamW updates (60 exposures, batch size 32), with weight decay 0.01, gradient clipping 1.0, and 30 warmup updates followed by a constant learning rate. Full train/validation evaluations occur at exposures 0, 10, 20, 40, and 60. Lowest development validation MAE selects one positive-exposure checkpoint per method/seed; ties favor the earlier checkpoint. All nine trajectories, all 36 fixed-budget snapshots, and all nine selections are hash-frozen before any new blind proposal, independently of performance. Fresh IID contains 400 independent cores (200 each at $t = 5, 7$); fresh OOD contains 800 (200 each at $t = 9, 11, 13, 15$). Development and previously proposed blind semantic hashes are excluded, and every route is evaluated on the same examples. No weights or test outcomes from the development search are transferred into this replication.

Complete fixed-plan analysis. The primary endpoint is fresh IID Raw-minus-MagicCore MAE for validation-selected checkpoints. OOD, Factorized comparisons, numerical spread, and common-budget views are secondary. Intervals resample paired training seeds and independent cores within fixed t strata, using 5,000 draws; secondary intervals are unadjusted. Tables 21 and 22 retain all common budgets and selections. Table 23 includes Raw, Factorized, zero, and train-mean comparisons. The selected-checkpoint prediction/target standard-deviation ratio is 0.926 on IID and 0.651 on OOD for MagicCore, compared with 0.107/0.111 for Raw and 0.330/0.336 for Factorized. Zero MAE is 0.2428/0.2329, and the training-mean reference gives 0.2427/0.2330 (IID/OOD).

B.4 STATE SUFFICIENCY AND READER ATTRIBUTION

This separate audit uses 960 training and 200 development joint-observable records at $t \in \{5, 7\}$, four active qubits, and 600 fresh-blind records balanced over $t \in \{9, 11, 13, 15\}$. All seven readers use 600 updates, batch size 32, and seeds 3407, 4519, and 8123; all 21 candidates are frozen before blind construction, independently of performance. Each reader receives the same final Pauli bits, coefficients, and joint-query tags. Explicit zero-state masks, coefficient-times-mask contributions, and component/final expectations are excluded. The column audit checks the exclusion of the explicit mask and contribution fields. Compatibility remains computable from the supplied bits as $m_b = \mathbf{1}[x_b = 0]$.

The isolated oracle reconstructs the target with maximum absolute error 5.6×10^{-17}. At the fixed 600-update budget, DeepSets explains held-out variance using these fields, whereas the tested Linear and pooled MLP readers have negative R^2. Relative to the relation-aware reader, the paired MAE gaps (comparator minus relation-aware) are 0.0187 for vanilla (95% CI $[-0.0036, 0.0451]$), 0.0113 for shuffled relations ($[-0.0087, 0.0365]$), and -0.0056 for sidecar-only ($[-0.0227, 0.0132]$). DeepSets is better by 0.0575 ($[0.0423, 0.0713]$), giving the strongest result among the seven readers under this

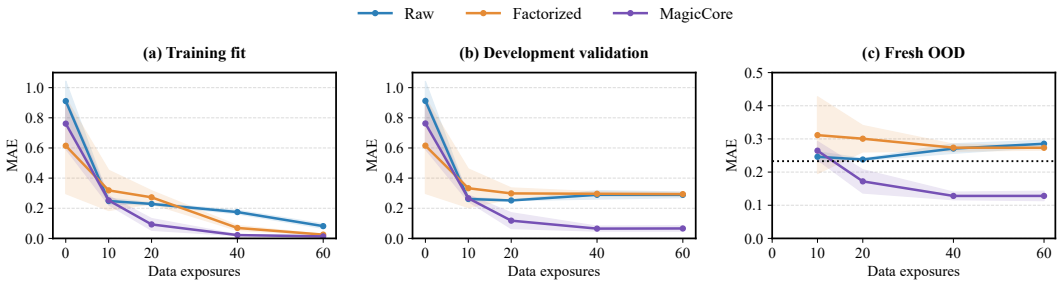


Figure 6: **Learning opportunity and generalization under the fixed confirmation plan.** Curves show the mean and shaded ± one sample standard deviation across all three new seeds. Train and development-validation MAE are shown at every scheduled evaluation, including initialization; fresh OOD shows all four frozen common-exposure checkpoints, not blind-selected models. The dotted line in (c) is the OOD zero reference. At 10 exposures MagicCore has not yet obtained an advantage; by 40–60 exposures its OOD MAE is approximately 0.128. Baseline training errors continue to decrease without a corresponding held-out improvement.

Table 21: **All four predeclared common-exposure checkpoints.** MAE mean ± sample standard deviation over all three seeds. Train and development validation are reused during optimization and selection; the two fresh sets are opened after the candidate freeze. **Best** and second best rank the learned routes within each exposure and column. No budget is selected using blind performance.

Exposures	Interface	Train	Dev. validation	Fresh IID	Fresh OOD
10	Raw	**0.2482 ± 0.0190**	**0.2617 ± 0.0129**	**0.2579 ± 0.0227**	**0.2462 ± 0.0150**
10	Factorized	0.3187 ± 0.1334	0.3329 ± 0.1278	0.3243 ± 0.1154	0.3114 ± 0.1158
10	MAGICCORE-LM	0.2523 ± 0.0298	0.2665 ± 0.0216	0.2620 ± 0.0164	0.2645 ± 0.0276
20	Raw	0.2288 ± 0.0076	0.2518 ± 0.0014	0.2432 ± 0.0021	0.2379 ± 0.0031
20	Factorized	0.2726 ± 0.0438	0.2985 ± 0.0362	0.2948 ± 0.0419	0.3007 ± 0.0396
20	MAGICCORE-LM	**0.0928 ± 0.0388**	**0.1178 ± 0.0524**	**0.1137 ± 0.0525**	**0.1720 ± 0.0351**
40	Raw	0.1747 ± 0.0125	0.2892 ± 0.0280	0.2773 ± 0.0147	0.2709 ± 0.0143
40	Factorized	0.0694 ± 0.0162	0.2961 ± 0.0133	0.2772 ± 0.0110	0.2736 ± 0.0105
40	MAGICCORE-LM	**0.0220 ± 0.0042**	**0.0649 ± 0.0162**	**0.0673 ± 0.0172**	**0.1281 ± 0.0115**
60	Raw	0.0814 ± 0.0148	0.2889 ± 0.0188	0.2914 ± 0.0052	0.2854 ± 0.0105
60	Factorized	0.0252 ± 0.0022	0.2937 ± 0.0116	0.2731 ± 0.0043	0.2733 ± 0.0070
60	MAGICCORE-LM	**0.0133 ± 0.0061**	**0.0660 ± 0.0177**	**0.0654 ± 0.0174**	**0.1281 ± 0.0141**

protocol. This audit compares neural readers with the exposed state held fixed; the raw/factorized experiments compare different input interfaces.

Shuffled relations use one deterministic permutation per joint-query pair: a CPU Torch generator is seeded by the first eight SHA-256 bytes of `relation-shuffle-v2|pair_id`, interpreted as an unsigned big-endian integer. The same permutation acts on both valid-token axes of every relation channel, while token features and query identities stay fixed and padding stays zero. This breaks token–relation alignment while preserving relation values and uses the same permutation at training and test. The audit's 0.94M-parameter sidecar is a standalone relation-pooling network (width 256, relation-weighted updates, mean/max pooling, and a query head), not the four-layer 1.964M frozen sidecar used in Table 7; the two parameter counts describe different models.

B.5 Active support, branch budgets, and computational cost

A separate frozen-reader diagnostic varies $k \in \{4, 6, 8, 10\}$ and $t \in \{2, 4, 6, 8, 10, 12, 14, 16\}$ independently at raw length $L = 256$, with six circuits and four joint queries per circuit in each cell. It evaluates the three-seed propagated-state audit Transformer and DeepSets of Appendix B.4, not the Compact readers in Table 1, on unconditioned random circuits beyond the balanced training family.

Table 22: **All validation-selected checkpoints by seed.** MAE is reported for train, development validation, and both fresh sets. Exposures are chosen independently for each method/seed from the same four-point grid; all trajectories nevertheless complete 60 exposures.

Seed	Interface	Exposures	Train	Dev. val.	Fresh IID	Fresh OOD
17389	Raw	20	0.2356	0.2534	0.2442	0.2412
17389	Factorized	60	0.0272	0.2948	0.2683	0.2702
17389	**MAGICCORE-LM**	60	0.0124	0.0523	0.0543	0.1183
29411	Raw	20	0.2205	0.2517	0.2446	0.2372
29411	Factorized	10	0.2304	0.2515	0.2440	0.2349
29411	**MAGICCORE-LM**	40	0.0188	0.0834	0.0871	0.1408
63127	Raw	20	0.2302	0.2505	0.2407	0.2352
63127	Factorized	10	0.2535	0.2669	0.2724	0.2546
63127	**MAGICCORE-LM**	40	0.0205	0.0581	0.0583	0.1249

Table 23: **Complete paired MAE comparisons for selected checkpoints.** Positive values favor MAGICCORE-LM. Percentile 95% intervals use 5,000 bootstrap draws of paired seeds and independent cores, stratified by t. The IID Raw comparison is primary; all other intervals are secondary and unadjusted. No seeds are excluded.

Test stratum	Comparator	Comparator − MagicCore	Paired 95% CI
IID	Raw	0.1766	[0.1556, 0.1958]
IID	Factorized	0.1950	[0.1597, 0.2243]
IID	Zero	0.1763	[0.1546, 0.1950]
IID	Train mean	0.1761	[0.1545, 0.1948]
OOD	Raw	0.1099	[0.0952, 0.1242]
OOD	Factorized	0.1252	[0.0957, 0.1520]
OOD	Zero	0.1049	[0.0915, 0.1172]
OOD	Train mean	0.1049	[0.0915, 0.1172]
$t = 9$	Raw	0.1528	[0.1295, 0.1743]
$t = 9$	Factorized	0.1692	[0.1354, 0.2002]
$t = 9$	Zero	0.1504	[0.1283, 0.1713]
$t = 9$	Train mean	0.1503	[0.1284, 0.1713]
$t = 11$	Raw	0.1226	[0.1019, 0.1441]
$t = 11$	Factorized	0.1418	[0.1037, 0.1819]
$t = 11$	Zero	0.1162	[0.0973, 0.1348]
$t = 11$	Train mean	0.1162	[0.0974, 0.1348]
$t = 13$	Raw	0.0945	[0.0756, 0.1130]
$t = 13$	Factorized	0.1082	[0.0767, 0.1403]
$t = 13$	Zero	0.0886	[0.0713, 0.1057]
$t = 13$	Train mean	0.0887	[0.0715, 0.1056]
$t = 15$	Raw	0.0695	[0.0542, 0.0857]
$t = 15$	Factorized	0.0816	[0.0541, 0.1119]
$t = 15$	Zero	0.0644	[0.0513, 0.0784]
$t = 15$	Train mean	0.0644	[0.0512, 0.0782]

Per-component branch budgets are $16, 32, 64, 128, 256, 512, 1024$ and untruncated; neural input is capped at 2048 joint tokens. Full-budget cells at $(k, t) = (6, 16), (8, 16), (10, 16)$ have incomplete coverage.

With increasing active support on this random-circuit distribution, targets concentrate near zero. Every fully covered untruncated cell at $k \geq 6$ has higher Transformer MAE than the zero reference; at $(k, t) = (10, 14)$ its MAE is 0.1716 versus 0.0055 for zero. These measurements characterize the

Table 24: **Same-input pre-contraction reader audit.** All learned readers receive the same exposed pre-contraction branch fields and are trained at $t \in \{5, 7\}$ before evaluation on 600 fresh continuous-target records balanced over $t \in \{9, 11, 13, 15\}$. Zero-state compatibility, coefficient-times-compatibility, component expectations, and the final target are withheld as explicit fields; compatibility is still computable from the supplied Pauli bits. Entries are mean ± sample standard deviation over three seeds. The analytic row is an isolated, non-learned exactness reference. All seven neural readers, prospectively specified and hash-frozen before blind generation, are shown; parameter matching applies to the 1.96M readers. Among learned readers, **green** marks the best mean and gray with underlining the second best; the analytic check is excluded from ranking.

Reader	Learned params	MAE ↓	R^2 ↑	Sign acc. (%) ↑
Linear	321	0.2351 ± 0.0006	−0.030 ± 0.003	50.1 ± 0.3
Pooled MLP	0.96M	0.2513 ± 0.0026	−0.266 ± 0.025	49.1 ± 0.7
DeepSets	1.96M	**0.1309 ± 0.0005**	**0.628 ± 0.002**	**87.1 ± 1.3**
Vanilla Transformer	1.96M	0.2070 ± 0.0286	0.145 ± 0.220	63.6 ± 10.1
Relation-aware Transformer	1.96M	0.1883 ± 0.0135	0.246 ± 0.063	69.8 ± 4.3
Shuffled relations	1.96M	0.1996 ± 0.0337	0.161 ± 0.205	66.2 ± 10.3
Relation sidecar only	0.94M	0.1827 ± 0.0054	0.318 ± 0.034	73.9 ± 2.2
Analytic contraction (audit only)	0	9.3×10^{-18}	1.000	100.0

Table 25: **Fresh-blind transfer to unseen circuit families.** Matched 1.96M-parameter Transformers train on random brickwork, star-entangler, and ring-cluster circuits at $t \in \{5, 7\}$, then evaluate 900 newly generated examples from three disjoint families. All nine candidates were frozen before the blind families were generated. Entries are mean ± sample standard deviation over three matched seeds. **Best** and second-best results. The aggregate paired MAE gain of MAGICCORE-LM is 0.3542 over Raw (95% CI $[0.3317, 0.3770]$) and 0.3426 over Factorized ($[0.3189, 0.3690]$).

	Family MAE ↓			Aggregate performance		
Interface	QAOA phase-gadget	Ladder excitation	Echo conjugation	MAE ↓	R^2 ↑	Sign (%) ↑
Raw	.3628 ± .0116	.6213 ± .0236	.3734 ± .0211	.4525 ± .0178	−.293 ± .109	51.6 ± 0.5
Factorized	.3584 ± .0207	.6032 ± .0104	.3609 ± .0086	.4408 ± .0091	−.211 ± .028	51.2 ± 1.9
MAGICCORE-LM	**.0517 ± .0082**	**.1766 ± .0194**	**.0665 ± .0163**	**.0983 ± .0145**	**.909 ± .021**	**95.0 ± 0.6**

interaction between active support, target scale, and reader generalization. Truncated analytic error and learned prediction error are recorded separately. Timing uses seed 3407 only and includes refactorization, propagation, tensorization, transfers, and synchronized inference; memory distinguishes total process allocation from added workspace. Repeated-query costs for $M \in \{1, 8, 32, 128\}$ reuse circuit preparation and rebuild branches for each observable, measuring amortization within this pipeline. Full grids and machine-readable measurements remain in the complexity–cost artifact.

B.6 CONTROLLED SYNTHETIC-FAMILY TRANSFER

The three routes train on random-brickwork, star-entangler, and ring-cluster circuits and are evaluated after freezing on QAOA phase-gadget, ladder-excitation, and echo-conjugation families. Across all three seeds, MAGICCORE-LM improves macro MAE over both matched routes. This earlier study

isolates a controlled synthetic-family shift under the propagated-state contract; Table 5 is the stronger main-text confirmation on recognizable fixed-angle algorithm templates.

B.7 PER-LABEL OOD CELLS

At primary seed 3407, frame-explicit factorized-to-MagicCore errors for negative/positive large targets fall from 0.6164/0.6148 to 0.5063/0.4942, while zero error rises from 0.0456 to 0.0904 and small-magnitude cells nearly tie. Across seven seeds, CNN/GNN exact-5 accuracies are 0.261/0.247, but calibrated posterior means reach the balanced zero-risk level, distinguishing coarse class signal from numerical prediction. Development-only checks, including temperature scaling (Guo et al., 2017), informed the shared frozen budget/calibration; aggregate BMAE remains this protocol's decision endpoint.

Table 26: **Primary-seed per-label OOD errors.** Primary-seed (3407) final blind OOD-BMAE cells, averaged across depths. The relation-aware, position-free reader mainly improves large-magnitude targets and slightly worsens zero/small targets. Within each target stratum, **green** marks the lower error and gray with underlining the second best.

	OOD error by target stratum ↓				
Method	**Zero**	**Neg.-large**	**Neg.-small**	**Pos.-small**	**Pos.-large**
Factorized-token	**.0456**	.6164	**.3581**	**.3589**	.6148
MAGICCORE-LM	.0904	**.5063**	.3618	.3641	**.4942**

B.8 FROZEN SIDECAR VERSUS SIDECAR–LLM ATTRIBUTION

Matched intervention and replay. This post-hoc audit reuses Table 7's 45 frozen candidates and the same ordered 2,000 IID and 2,000 depth-512 OOD examples. The only intervention is to remove the adapted LLM residual in Equation 5, retaining the route's frozen sidecar, input construction, five class centers, and temperature-one posterior. No retraining, calibration, selection, or new test generation occurs. One sidecar per route is shared across all five backbones and three adaptation seeds; sidecar-only therefore has no adaptation-seed SD. We run each sidecar directly in FP32 with the original batch size two, verify its predictions against the sidecar posterior recovered from cached fusion probabilities/residual logits, and reproduce the original per-seed and aggregate metrics. Maximum scalar replay disagreement is below 2.3×10^{-7}. Checkpoint and ordered-example identities are verified before analysis.

Paired uncertainty. Define $\Delta = \mathrm{BMAE}(\text{sidecar}) - \mathrm{BMAE}(\text{sidecar+LLM})$, positive for an LLM-path gain. The 10,000 crossed bootstrap draws (RNG seed 20260915) resample three paired adaptation seeds and independent test parents within target-label/rewrite strata. The same draws apply across routes and backbones. Macro results average the fixed five-backbone panel; backbones are not resampled or counted as additional training seeds. Intervals are pointwise, unadjusted, and conditional on the three fixed sidecars. Tables 27 and 28 retain every route/backbone; Table 29 reports all seed differences. Test-parent-only intervals and additional metrics are archived.

Incremental LLM-path effect. For MagicCore, macro OOD BMAE falls from 0.398785 to 0.395477: an absolute gain of 0.003309 (95% CI $[0.002017, 0.004886]$), or 0.83%. All five backbone mean gains are positive, with positive intervals for Qwen, Gemma, and Phi and intervals crossing zero for DeepSeek and Llama. Thirteen of the fifteen adaptation-seed OOD differences are positive. The IID macro gain is 0.002773 ($[0.001570, 0.004194]$). This estimates the added effect of the complete adapted LLM path on these frozen sidecars; Table 7's Raw/Factorized comparisons instead measure the benefit of choosing the MagicCore interface within each LLM pipeline.

Table 27: **Frozen sidecar ablation: IID.** Same models and examples as Table 7; lower BMAE is better. Sidecar+LLM is mean ± SD across three adaptation seeds. Positive Δ favors adding the LLM path; intervals are paired, pointwise 95% CIs. Within each row, **green** marks lower mean error and gray with underlining the higher mean, not statistical significance.

Interface	Sidecar	Sidecar+LLM	Δ	95% CI
Qwen3-4B				
Raw	0.492268	0.482034±0.000136	+0.010234	[+0.008331, +0.012128]
Factorized	0.456636	0.449183±0.001781	+0.007453	[+0.005400, +0.009587]
MagicCore	0.392975	0.389048±0.002159	+0.003926	[+0.001467, +0.006054]
Gemma-3-4B				
Raw	0.492268	0.483254±0.001601	+0.009014	[+0.006443, +0.011572]
Factorized	0.456636	0.451416±0.002055	+0.005220	[+0.002801, +0.007319]
MagicCore	0.392975	0.390758±0.002171	+0.002216	[+0.000229, +0.004568]
Phi-4-mini				
Raw	0.492268	0.482433±0.000185	+0.009835	[+0.007591, +0.012050]
Factorized	0.456636	0.448261±0.000839	+0.008375	[+0.006734, +0.010003]
MagicCore	0.392975	0.390354±0.000452	+0.002621	[+0.001349, +0.003901]
DeepSeek-R1-Qwen-7B				
Raw	0.492268	0.482863±0.000865	+0.009405	[+0.007571, +0.011187]
Factorized	0.456636	0.449109±0.000924	+0.007528	[+0.005976, +0.009131]
MagicCore	0.392975	0.389831±0.002727	+0.003144	[$+1.0\times10^{-7}$, +0.005542]
Llama-3.1-8B				
Raw	0.492268	0.482135±0.000350	+0.010133	[+0.007605, +0.012725]
Factorized	0.456636	0.454826±0.001600	+0.001810	[+0.000342, +0.003948]
MagicCore	0.392975	0.391014±0.002080	+0.001961	[−0.000362, +0.004178]
Fixed-five macro				
Raw	0.492268	0.482544±0.000321	+0.009724	[+0.007836, +0.011649]
Factorized	0.456636	0.450559±0.000399	+0.006077	[+0.004945, +0.007242]
MagicCore	0.392975	0.390201±0.001084	+0.002773	[+0.001570, +0.004194]

Table 28: **Frozen sidecar ablation: depth-512 OOD.** Same models and examples as Table 7; lower BMAE is better. Sidecar+LLM is mean ± SD across three adaptation seeds. Positive Δ favors adding the LLM path; intervals are paired, pointwise 95% CIs. Within each row, **green** marks lower mean error and gray with underlining the higher mean, not statistical significance.

Interface	Sidecar	Sidecar+LLM	Δ	95% CI
Qwen3-4B				
Raw	0.485117	0.471933±0.002436	+0.013185	[+0.010808, +0.016278]
Factorized	0.441528	0.434499±0.001660	+0.007029	[+0.005060, +0.009000]
MagicCore	0.398785	0.394432±0.002493	+0.004353	[+0.001567, +0.006714]
Gemma-3-4B				
Raw	0.485117	0.474390±0.004576	+0.010727	[+0.005813, +0.015172]
Factorized	0.441528	0.436325±0.002260	+0.005203	[+0.002551, +0.007436]
MagicCore	0.398785	0.395651±0.002791	+0.003134	[+0.000840, +0.006078]
Phi-4-mini				
Raw	0.485117	0.473094±0.003370	+0.012024	[+0.008162, +0.015433]
Factorized	0.441528	0.433783±0.000833	+0.007744	[+0.006153, +0.009310]
MagicCore	0.398785	0.395705±0.000645	+0.003080	[+0.001684, +0.004502]
DeepSeek-R1-Qwen-7B				
Raw	0.485117	0.473193±0.000482	+0.011925	[+0.010534, +0.013477]
Factorized	0.441528	0.434238±0.000449	+0.007289	[+0.005939, +0.008644]
MagicCore	0.398785	0.394957±0.003327	+0.003828	[−0.000004, +0.006602]
Llama-3.1-8B				
Raw	0.485117	0.471264±0.002020	+0.013853	[+0.011161, +0.016866]
Factorized	0.441528	0.439907±0.000929	+0.001621	[+0.000503, +0.003120]
MagicCore	0.398785	0.396638±0.002236	+0.002148	[−0.000318, +0.004474]
Fixed-five macro				
Raw	0.485117	0.472775±0.001598	+0.012343	[+0.010256, +0.014700]
Factorized	0.441528	0.435751±0.000410	+0.005777	[+0.004638, +0.006897]
MagicCore	0.398785	0.395477±0.001260	+0.003309	[+0.002017, +0.004886]

Table 29: **Adaptation-seed sensitivity of the frozen sidecar ablation.** Each entry is sidecar-only minus sidecar+LLM BMAE, positive for an LLM-path gain. All three adaptation seeds share one frozen sidecar per route. The very small DeepSeek IID MagicCore gain at seed 3408 is within numerical replay tolerance.

	IID gain			Depth-512 OOD gain		
Interface	3407	3408	3409	3407	3408	3409
Qwen3-4B						
Raw	+0.010333	+0.010290	+0.010080	+0.015994	+0.011899	+0.011661
Factorized	+0.007549	+0.005626	+0.009185	+0.007447	+0.005199	+0.008439
MagicCore	+0.001434	+0.005230	+0.005114	+0.001474	+0.005842	+0.005743
Gemma-3-4B						
Raw	+0.009815	+0.007171	+0.010057	+0.011503	+0.005813	+0.014865
Factorized	+0.006306	+0.006504	+0.002851	+0.006045	+0.006921	+0.002642
MagicCore	+0.000283	+0.001801	+0.004564	+0.000967	+0.002154	+0.006283
Phi-4-mini						
Raw	+0.009938	+0.009622	+0.009945	+0.014485	+0.013403	+0.008183
Factorized	+0.008434	+0.009183	+0.007509	+0.007473	+0.008679	+0.007081
MagicCore	+0.002151	+0.002658	+0.003053	+0.002573	+0.003807	+0.002861
DeepSeek-R1-Qwen-7B						
Raw	+0.009674	+0.010103	+0.008437	+0.012481	+0.011636	+0.011657
Factorized	+0.006662	+0.007420	+0.008501	+0.006870	+0.007235	+0.007764
MagicCore	+0.004880	$+4.6\times10^{-8}$	+0.004550	+0.005566	−0.000008	+0.005927
Llama-3.1-8B						
Raw	+0.010344	+0.009729	+0.010325	+0.015896	+0.011856	+0.013808
Factorized	+0.003636	+0.001145	+0.000650	+0.002426	+0.001832	+0.000605
MagicCore	+0.003653	−0.000361	+0.002589	+0.004001	−0.000335	+0.002777
Fixed-five macro						
Raw	+0.010021	+0.009383	+0.009769	+0.014072	+0.010921	+0.012035
Factorized	+0.006517	+0.005976	+0.005739	+0.006052	+0.005973	+0.005306
MagicCore	+0.002480	+0.001866	+0.003974	+0.002916	+0.002292	+0.004718