Universal Quantum Transformer

Sungyong Chung[1] and Alireza Talebpour[*1]

[1] *Grainger College of Engineering, Department of Civil and Environmental Engineering, University of Illinois Urbana-Champaign*

Abstract

Classical continuous-space neural networks fundamentally struggle to lock into exact mathematical symmetries, such as modular arithmetic and non-commutative algebra. To approximate these discrete logical rules, they often rely on massive parameter scaling, resulting in stochastic instability even after delayed generalization phenomena known as grokking. Here, we introduce the Universal Quantum Transformer (UQT), a fundamentally novel, quantum-native computing architecture that uses the physical properties of multi-qubit systems as a universal inductive bias for exact mathematical and algebraic reasoning. Rather than translating classical neural mechanisms, our framework relies entirely on parameterized geometric phase embedding and $SU(2)$ wave-interference. We demonstrate that the quantum attention circuit, operating on a highly compact 5-qubit substrate, perfectly learns two highly distinct formal classes: cyclic modular arithmetic ($\mathbb{Z}_{11}$) and non-Abelian algebra (the S_4 permutation group). While classical attention-based networks exhibit stochastic instability at convergence, the UQT achieves mathematically exact, deterministic generalization. We refer to this phenomenon as crystallization: a step beyond the well-known phenomenon of grokking. Crucially, this framework yields massive computational and memory advantages by theoretically bypassing the quadratic bottleneck of classical self-attention, and by logarithmically compressing the required representation dimension to eliminate the massive over-parameterization inherent to classical networks. Finally, we deploy this architecture on noisy intermediate-scale quantum (NISQ) hardware, proving its viability on current IBM Quantum computers. These results establish parameterized quantum topology as a universally superior physical substrate for exact artificial intelligence.

While attention-based models have achieved remarkable results in natural language processing, their capacity for systematic mathematical reasoning across cyclic arithmetic and non-commutative algebra remains brittle, with models often failing to generalize on structurally identical problems outside their training distribution [1, 2]. Standard architectures operating in unconstrained continuous Euclidean space ($\mathbb{R}^n$) carry inductive biases misaligned with discrete, periodic, and non-commutative algebraic structure, not because such representations are theoretically impossible, but because they are not privileged by design. To solve exact arithmetic, classical models are forced to rely on massive parameter scaling, attempting to brute-force discrete geometric rules into high-dimensional vector spaces. As recent studies on algorithmic grokking reveal, this Euclidean approach causes networks to default to pure memorization, struggling to generalize to unseen operands without extensive, delayed optimization [3, 4]. Yet, even when classical networks finally achieve this delayed generalization, they still merely approximate the discrete mathematics. Fundamentally, theoretical analyses show that classical continuous-space models must rely on massive over-parameterization to statistically simulate wave interference [5]. As we formally establish in the following section, this continuous approximation fundamentally leaves classical attention-based architectures vulnerable to persistent stochastic instability at convergence and erratic predictions on unseen operands.

[*]Corresponding author: ataleb@illinois.edu.

We resolve this limitation by introducing the Universal Quantum Transformer (UQT). Recent efforts to develop quantum Transformers have primarily focused on translating classical neural mechanisms, such as attention matrices and softmax functions, into parameterized quantum circuits [6, 7] or fault-tolerant linear algebra routines [8, 9]. Unfortunately, because these approaches structurally mirror classical continuous-space architectures, they inherit the same misaligned inductive biases, forcing them to rely on continuous approximations rather than natively resolving discrete logic. Unlike the previous efforts, the UQT is not a quantum analogue of a classical neural network; rather, it is a completely novel computational methodology that maps abstract symbols directly onto the unitary operations of a parameterized quantum system. By leveraging the inherent periodicity of quantum phase and the non-commutative geometry of $SU(2)$ rotations, our architecture provides a universal physical inductive bias for exact mathematical reasoning. In this study, we establish that a single, unified topological modality, i.e., the quantum attention circuit, jointly embeds tokens into a shared superposition to natively resolve exact arithmetic and non-commutative algebra through global wave-interference, completely bypassing the need for massive over-parameterization.

FROM GROKKING TO CRYSTALLIZATION

To contextualize the physical advantage of the UQT, it is necessary to formally distinguish between the transient learning dynamics of algorithmic generalization and the structural stability of a model's converged representation. In the machine learning literature, the delayed discovery of generalizing solutions is broadly referred to as grokking [3, 4].

Definition 1: Grokking. *Let t denote the optimization epoch, and let $\mathcal{A}_{\text{train}}(t)$ and $\mathcal{A}_{\text{test}}(t)$ denote the training and test accuracies at epoch t, respectively. Let t_{mem} be the epoch at which the model perfectly memorizes the training dataset, i.e., $\mathcal{A}_{\text{train}}(t_{\text{mem}}) = 100\%$. Grokking is the phenomenon in which $\mathcal{A}_{\text{test}}(t)$ remains near random chance until a much later epoch $t_{\text{grok}} \gg t_{\text{mem}}$, after which it rapidly increases toward* 100%.

However, because prior work has often treated grokking as a single, monolithic transition in generalization, it does not distinguish between approximate statistical generalization and exact recovery of the underlying mathematical rule. Achieving grokking merely guarantees that a neural network has found a statistical decision boundary that adequately covers the unseen data. When classical continuous-space Transformers learn algorithmic tasks, their generalization often emerges only after an extended memorization phase, and the resulting solution may remain an approximate statistical representation rather than an exact symbolic or algebraic resolution. Stochastic gradient updates induced by mini-batch training can perturb the continuous parameters away from the narrow alignments needed to approximate discrete logical structure. Consequently, even when classical models exhibit grokking on the training distribution, their generalization may remain statistically fragile, as reflected by persistent accuracy oscillations and nonzero test-accuracy variance, i.e., $\text{Var}[\mathcal{A}_{\text{test}}(t)] > 0$, as we show in the following sections. To capture the ultimate resolution of formal mathematical systems, we introduce the stricter concept of crystallization.

Definition 2: Crystallization. *A strict regime of generalization that requires, but supersedes, grokking. A network undergoes crystallization at epoch $t_c \geq t_{\text{grok}}$ if $\mathcal{A}_{\text{test}}(t_c) = 100\%$ and the model achieves deterministic stability, resulting in strictly zero variance for all subsequent epochs:* $\text{Var}[\mathcal{A}_{\text{test}}(t)] = 0$ *for all* $t \geq t_c$.

While highly restricted classical toy models can be mathematically engineered to crystallize on simple commutative arithmetic via massive over-parameterization [5], standard classical Transformers operating in unconstrained Euclidean space ($\mathbb{R}^n$) fundamentally struggle to achieve this zero-variance state. Unlike the statistical approximation inherent to classical

grokking, crystallization guarantees that the network has physically embodied the target logic.

QUANTUM ATTENTION CIRCUITS

To enable a truly global evaluation of mathematical symmetries, our proposed architecture leverages the compositionality of quantum operators. Specifically, in the quantum attention circuit (Figure 1a), a sequence of S tokens $(x_1, \ldots, x_S)$ is mapped through an embedding table to a set of physical rotation angles that parameterize unitary transformations on an N_{emb}-qubit subspace of the total N-qubit register. These parameterized unitaries are applied consecutively (from $U(\vec{\theta}_{x_1})$ to $U(\vec{\theta}_{x_S})$) entirely prior to any explicit logical mixing operations. Due to the intrinsic compositional structure of quantum mechanics, where unitary operators combine via matrix multiplication, this sequence does not merely stack independent transformations but produces a joint operation that encodes the interaction between operands directly in the geometry of the resulting quantum state. The multi-qubit wave function therefore acts as a native geometric calculator, in which the relative phases introduced by each operand interfere constructively or destructively. The resulting superposed state thus contains a distributed encoding of operand interactions, which is subsequently processed by L deeper mixing layers to decode and extract task-relevant features.

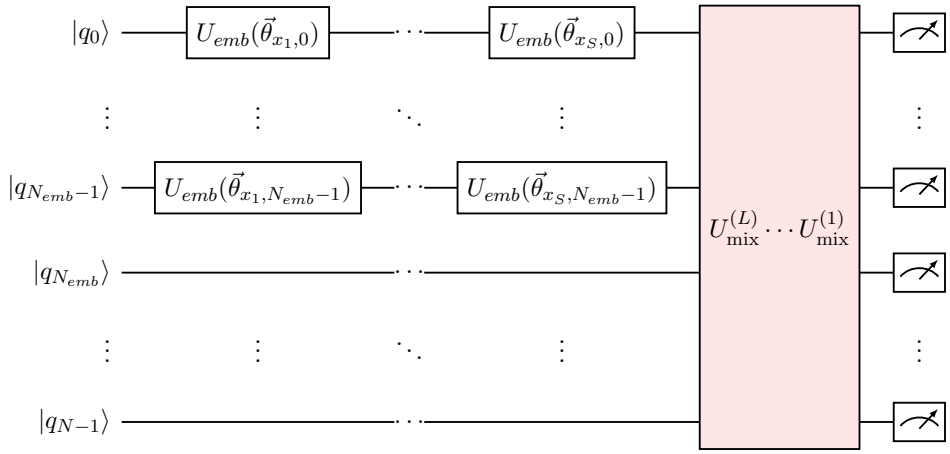
As detailed in Figure 1b, a single mixing layer $U_{\text{mix}}^{(l)}$ consists of parameterized single-qubit general $SU(2)$ Euler rotations $(U_{\text{Rot}}^{(l)})$ applied to all N qubits, followed by a global, cyclic conveyor-belt ring of CNOT gates. Because physical noisy intermediate-scale quantum (NISQ) hardware suffers severe decoherence when executing complex multi-qubit entangling operations, this hardware-efficient topology generates deep global entanglement using only fundamental 2-qubit CNOT connections, providing a logical backbone capable of natively solving formal mathematical systems.

In this formulation, representation (phase-based embedding) is cleanly decoupled from reasoning (quantum interference followed by decoding), enabling a unified framework in which algebraic structure is evaluated through the native physics of the system.

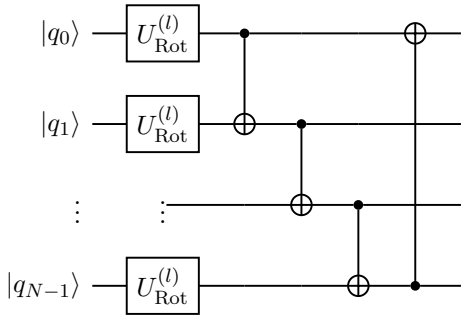
Across our empirical evaluations, we instantiated the UQT on a highly compact $N = 5$ qubit register. The total parameter capacity of the architecture is exceptionally compact, defined strictly by the sum of the token embedding rotations and the parameterized entanglement logic gates. For a token space of size V embedded onto a subspace of N_{emb} qubits (utilizing 4 rotations per qubit), and L total mixing layers across all N qubits, the parameter count is exactly $(V \times N_{emb} \times 4) + (L \times N \times 3)$. Unlike classical continuous-space models that require hundreds of thousands of weights to approximate discrete logic [3, 5], the UQT isolates the exact algebraic geometries using a footprint of only a few hundred parameters.

CRYSTALLIZATION IN MODULAR ARITHMETIC

We tasked the quantum attention circuit with learning modular arithmetic given two input tokens, x_1 and x_2. To rigorously evaluate its geometric learning capabilities, we tested the architecture across three distinct algebraic regimes. First, we evaluated addition over the full prime Galois field $\mathbb{Z}_{11}$, where the network must predict $x_1 + x_2 \pmod{11}$ (Figures 2a and 2b). Second, we evaluated multiplication over the full field $\mathbb{Z}_{11}$ ($x_1 \times x_2 \pmod{11}$), which includes the zero element (Figures 2c and 2d). Finally, we evaluated multiplication restricted strictly to the bijective multiplicative group $\mathbb{Z}_{11}^* = \{1, 2, \ldots, 10\}$, which excludes the zero element (Figures 2e and 2f). In our implementation, we mapped the token space across an $N_{emb} = 4$ qubit subspace and utilized a mixing depth of $L = 25$. Measuring the full 5-qubit register ($N = 5$), this architecture requires only 551 trainable parameters.



(a) Quantum attention circuit



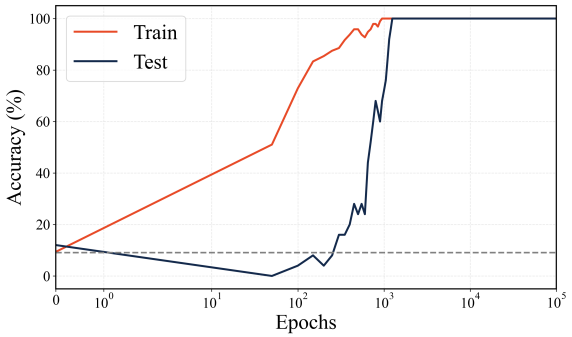
(b) Single mixing layer architecture ($U^{(l)}_{\text{mix}}$)

Figure 1: **Topology of the Universal Quantum Transformer (UQT).** The architecture strictly generalizes to arbitrary N-qubit registers and sequence lengths (S). **(a)** In the quantum attention circuit, S sequential tokens $(x_1, \ldots, x_S)$ are embedded entirely prior to structural entanglement. The superposed phases are processed simultaneously by the L-layer mixing block. Measurements are performed on all N qubits to generate the output. **(b)** A single layer l of the shared mixing sequence, $U^{(l)}_{\text{mix}}$, consisting of parameterized single-qubit general $SU(2)$ Euler gates ($U^{(l)}_{\text{Rot}}$) and a global cyclic conveyor-belt ring of CNOT gates.

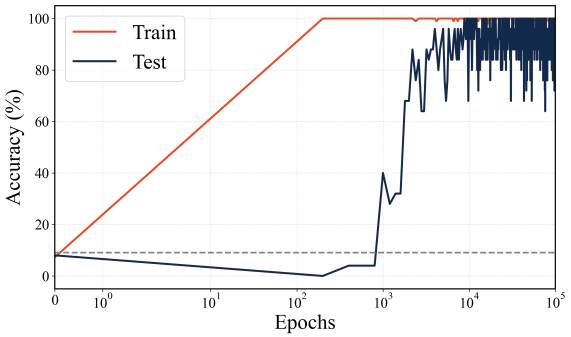
Modular arithmetic represents a rigid test of machine learning, as the cyclic algebraic structure defies continuous linear approximation. Consequently, while classical models trained on such tasks often exhibit grokking [3], they fundamentally struggle to crystallize. To contextualize the physical advantage of our architecture, we trained classical attention-based Transformers [10] as a baseline.

As observed in Figures 2b, 2d, and 2f, the classical Transformer requires massive overparameterization to brute-force the cyclic geometry into a high-dimensional Euclidean space. Although the classical models successfully grok the dataset, exhibiting a delayed transition in test accuracy long after reaching 100% training accuracy, they fail to maintain stable generalization. Instead, the test accuracy exhibits severe, erratic fluctuations. Furthermore, the classical network relies heavily on statistical pattern matching. For instance, it easily isolates and memorizes the simple zero-rule ($x_1 \times 0 = 0 \pmod{11}$) during early epochs of training. This classical shortcut artificially raises the lower bound of the test accuracy, buffering the lowest dips of the oscillations to remain above 80% (Figure 2d). Once the zero element is excluded, the test accuracy oscillates violently between 70% and 100% (Figure 2f).

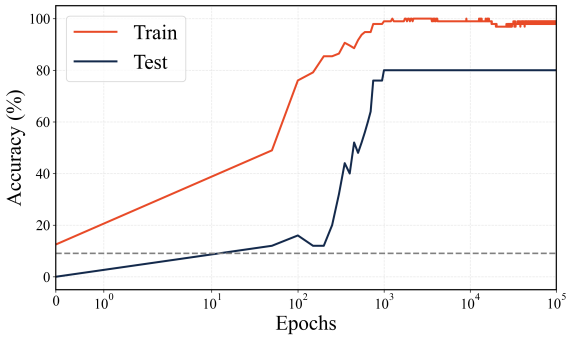
In stark contrast to the classical model, the UQT is fundamentally incapable of taking



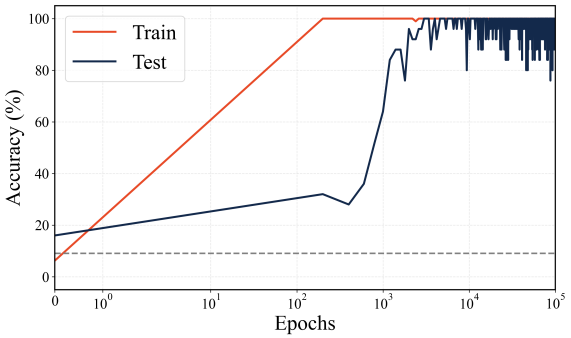
(a) UQT: Mod 11 addition ($\mathbb{Z}_{11}$)



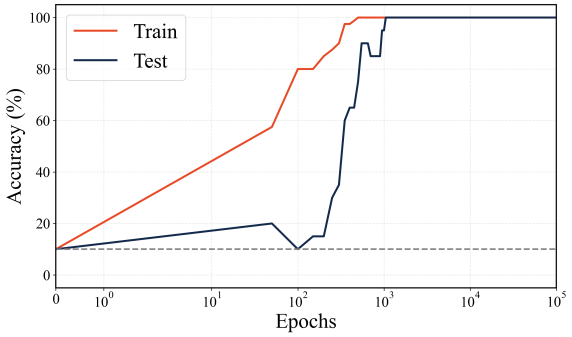
(b) Classical: Mod 11 addition ($\mathbb{Z}_{11}$)



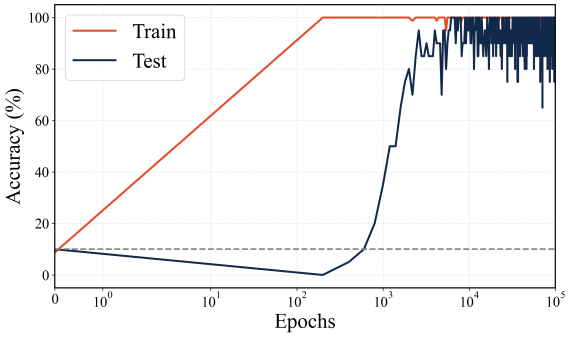
(c) UQT: Mod 11 multiplication ($\mathbb{Z}_{11}$)



(d) Classical: Mod 11 multiplication ($\mathbb{Z}_{11}$)



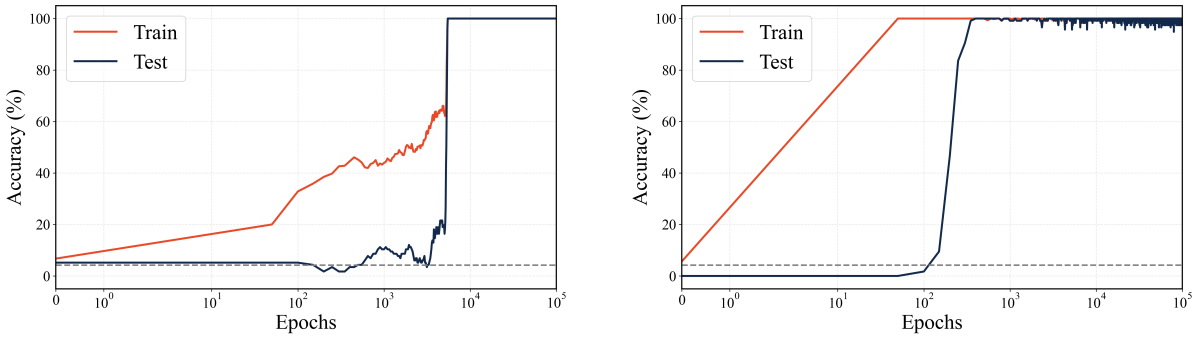
(e) UQT: Mod 11 multiplication ($\mathbb{Z}_{11}^*$)



(f) Classical: Mod 11 multiplication ($\mathbb{Z}_{11}^*$)

Figure 2: **Quantum crystallization in modular arithmetic. (a, b)** Both architectures learn addition, but the classical transformer exhibits severe stochastic instability at convergence. **(c)** The UQT physically fails to learn multiplication with zero, as the irreversible many-to-one mapping ($x_1 \times 0 = 0$) violates the unitary constraints of quantum mechanics ($U^\dagger U = I$). **(d)** The classical transformer memorizes the zero-rule via pattern matching, artificially buffering its instability. **(e, f)** When restricted to the bijective $\mathbb{Z}_{11}^*$ multiplicative group (excluding the zero element), the UQT achieves exact, deterministic crystallization, whereas the classical network oscillates violently.

the zero-rule shortcut. As demonstrated in Figure 2c, when the UQT attempts to learn the complete $\mathbb{Z}_{11}$ multiplication field (including the 0 element), the network fails to reach 100% accuracy. This failure is a direct consequence of pure quantum mechanics. Multiplication by zero is a many-to-one mapping that inherently destroys information, rendering the operation fundamentally irreversible. Because quantum evolution is strictly governed by unitary operators ($U^\dagger U = I$), the network cannot natively execute an irreversible mapping without collapsing



(a) UQT: S_4 permutations

(b) Classical: S_4 permutations

Figure 3: **Quantum crystallization in non-Abelian algebra.** The non-commutative geometry of $SU(2)$ allows the quantum model to cleanly lock into the S_4 group laws. In contrast, the classical Transformer struggles to maintain stable generalization due to its continuous Euclidean geometry.

its own topological geometry. Addition, however, remains perfectly bijective even with zero ($x_1 + 0 = x_1$), allowing the UQT to crystallize perfectly (Figure 2a).

When evaluated on the restricted $\mathbb{Z}_{11}^*$ multiplicative group (Figure 2e), the zero element is eliminated. Within this regime, every multiplication operation becomes a perfect bijection, a reversible cyclic permutation of the set. Restored to a purely unitary framework, the UQT perfectly generalizes on the dataset, crystallizing into a deterministic 100% accuracy without the stochastic instability characteristic of classical networks.

CRYSTALLIZATION IN NON-ABELIAN ALGEBRA

While addition and multiplication are commutative, spatial and physical transformations are inherently non-commutative. To test this regime, we evaluated the architecture on the symmetric permutation group S_4. This abstract group consists of 24 distinct elements, meaning the network must learn its entire Cayley table. Here, a Cayley table is a strict algebraic multiplication table, defining the exact outcomes of all $24 \times 24 = 576$ possible paired combinations. Because S_4 is non-Abelian (i.e., applying permutation A followed by B yields a fundamentally different state than B followed by A), classical networks operating in flat, commutative Euclidean space frequently struggle to learn these operations efficiently [4]. To approximate these discrete, non-commutative rules, classical models must rely on massive over-parameterization. Consequently, even when these networks eventually grok the dataset, they fail to crystallize, resulting in the severe stochastic instability observed at convergence (Figure 3b).

The quantum architecture, however, provides a native algebraic solution. The UQT embeds tokens using $SU(2)$ rotation matrices, the fundamental mathematical group that governs quantum spin and 3D spatial rotations. Because $SU(2)$ operations are intrinsically non-commutative, the physical wave-interference of the sequential quantum gates perfectly mirrors the geometric rules of the target algebra. By mapping the 24 abstract permutations directly into this continuous, non-commutative phase space, the UQT successfully crystallized across all 576 equations without statistical drift (Figure 3a). Note that to accommodate the higher representational density of the 24 permutations ($V = 24$), we expanded the embeddings to the full register ($N_{emb} = 5$) and reduced the mixing depth to $L = 14$. Measuring the full 5-qubit register ($N = 5$), this model requires a total of 690 parameters.

PHYSICAL DEPLOYMENT ON NISQ HARDWARE

To empirically validate that the generalization capabilities of the UQT are not merely artifacts of noiseless classical simulation, we deployed the architecture onto physical superconducting quantum processors. In classical deep learning, gradients are computed via backpropagation. Physical quantum processors, however, lack an analytic computational graph, meaning exact analytic gradients must be evaluated empirically using the Parameter-Shift Rule [11, 12]. While the UQT's extreme parameter efficiency (less than 700 parameters for these computational tasks) makes physical training theoretically viable, current NISQ devices suffer from two-qubit gate infidelities. The deep unitary depth required for structural entanglement ($L = 25$ layers in arithmetic tasks, necessitating > 100 sequential CNOT gates per forward pass) places unmitigated gradient-evaluation loops beyond the fidelity horizon of bare superconducting qubits.

To bridge this gap, we employed an offline training protocol: the UQT was trained via noiseless JAX JIT-compiled automatic differentiation to achieve the algorithmic phase transition. The trained serialized parameters ($\vec{\theta}_{opt}$) were then compiled into static physical gates and evaluated across the tasks on IBM Quantum hardware.

As detailed in Table 1, the quantum architecture demonstrated robust physical resilience. Across 30 unmitigated hardware evaluations, the UQT achieved an overall physical success rate of 96.7% (29/30). It is crucial to interpret the raw confidence values through the lens of quantum measurement dimensionality. The Hilbert space for these 5-qubit evaluations encompasses $2^5 = 32$ basis states, resulting in a random physical noise floor of $\sim 3.1\%$. The recorded confidence peaks across the passing evaluations (7% to 24%) represent significant, statistically decisive constructive interference overriding environmental decoherence.

The single misclassification occurred within the training set of the highly complex S_4 non-Abelian domain, where the network erroneously predicted Class 13 instead of the target Class 18. However, analysis of the raw probability distribution generated across 8,192 shots (Figure 4) revealed that the true target remained the distinct second-most probable state (4.7%), successfully maintaining a probability peak above the 3.1% random noise floor. This specific failure mode illustrates the boundary of current NISQ fidelity; while the geometric wave-interference correctly isolated the target amplitude within the simulated model, unmitigated hardware decoherence allowed an erroneous state ($|13\rangle$) to marginally overtake the primary signal (6.3%). Despite this, the 100% success rate across all unseen generalization test sets confirms that the UQT's geometric parameterization inherently generates highly stable physical wave-interference patterns capable of surviving modern NISQ constraints.

ASYMPTOTIC COMPLEXITY

Consider the representational scaling required to model formal mathematical systems. In standard classical Transformers, embedding a finite token space of size V requires mapping tokens into a high-dimensional Euclidean space, with an embedding table demanding $\mathcal{O}(V \times d_{\text{model}})$ parameters. Furthermore, classical multi-head attention and feed-forward networks rely on dense projection matrices that scale quadratically with this hidden dimension, imposing a massive $\mathcal{O}(d_{\text{model}}^2)$ parameter bottleneck. Conversely, the UQT maps tokens directly into the continuous phase amplitudes of an N-qubit register. Because the multi-qubit Hilbert space scales exponentially ($\mathcal{O}(2^N)$ state dimensions), the UQT achieves massive representational density using only a logarithmic number of physical qubits: $\mathcal{O}(\log V)$. Consequently, the quantum embedding dimension is strictly bounded by the number of phase angles applied to the N qubits, reducing the representation scaling to $\mathcal{O}(V \times \log V)$, while the sequential quantum mixing layers require only $\mathcal{O}(L \times \log V)$ parameters. This allows the UQT to bypass the $\mathcal{O}(d_{\text{model}}^2)$ classical parameter explosion entirely.

Furthermore, the dense, all-to-all classical softmax attention matrix requires $\mathcal{O}(S^2 \cdot d_{\text{model}})$

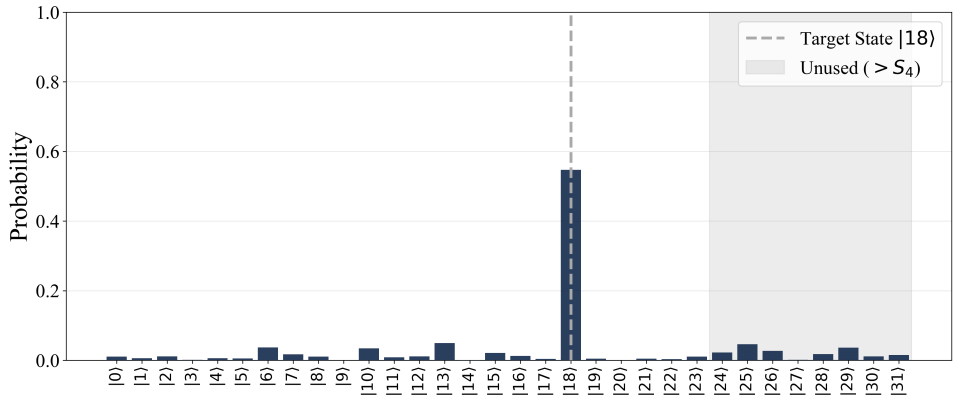
Table 1: **NISQ hardware inference results.** Unmitigated evaluation of the trained UQT parameters on physical IBM Quantum processors. The architecture achieved an overall 96.7% success rate (29/30), including 100% accuracy on all unseen generalization test sets. Executions were distributed across distinct hardware systems (`ibm_marrakesh` and `ibm_fez`) based on queue availability.

Domains	Input / Sequence	Target	IBM Output	Confidence	Status
Domain 1: $\mathbb{Z}_{11}$ Modular Addition (`ibm_marrakesh`)					
Train (Memorization)	$9+4 \pmod{11}$	2	2	22.0%	**PASS**
	$1+10 \pmod{11}$	0	0	22.4%	**PASS**
	$2+7 \pmod{11}$	9	9	14.7%	**PASS**
	$4+7 \pmod{11}$	0	0	24.1%	**PASS**
	$6+6 \pmod{11}$	1	1	19.1%	**PASS**
Test (Generalization)	$5+0 \pmod{11}$	5	5	20.5%	**PASS**
	$3+3 \pmod{11}$	6	6	19.1%	**PASS**
	$10+8 \pmod{11}$	7	7	17.6%	**PASS**
	$5+10 \pmod{11}$	4	4	13.9%	**PASS**
	$3+9 \pmod{11}$	1	1	19.4%	**PASS**
Domain 2: $\mathbb{Z}_{11}^*$ Modular Multiplication (`ibm_fez`)					
Train (Memorization)	$3\times5 \pmod{11}$	4	4	12.1%	**PASS**
	$6\times6 \pmod{11}$	3	3	11.5%	**PASS**
	$1\times8 \pmod{11}$	8	8	10.8%	**PASS**
	$1\times4 \pmod{11}$	4	4	8.5%	**PASS**
	$4\times6 \pmod{11}$	2	2	8.0%	**PASS**
Test (Generalization)	$6\times4 \pmod{11}$	2	2	7.7%	**PASS**
	$5\times5 \pmod{11}$	3	3	9.5%	**PASS**
	$4\times10 \pmod{11}$	7	7	13.1%	**PASS**
	$5\times6 \pmod{11}$	8	8	10.4%	**PASS**
	$2\times9 \pmod{11}$	7	7	7.8%	**PASS**
Domain 3: S_4 Non-Abelian Permutations (`ibm_marrakesh`)					
Train (Memorization)	$P(3)\circ P(22)$	10	10	9.7%	**PASS**
	$P(15)\circ P(17)$	18	13	6.3%	**FAIL***
	$P(19)\circ P(1)$	18	18	14.2%	**PASS**
	$P(7)\circ P(23)$	16	16	9.1%	**PASS**
	$P(6)\circ P(13)$	15	15	13.5%	**PASS**
Test (Generalization)	$P(9)\circ P(19)$	1	1	13.8%	**PASS**
	$P(23)\circ P(11)$	12	12	9.5%	**PASS**
	$P(4)\circ P(13)$	6	6	10.6%	**PASS**
	$P(3)\circ P(5)$	1	1	9.9%	**PASS**
	$P(18)\circ P(8)$	1	1	9.3%	**PASS**

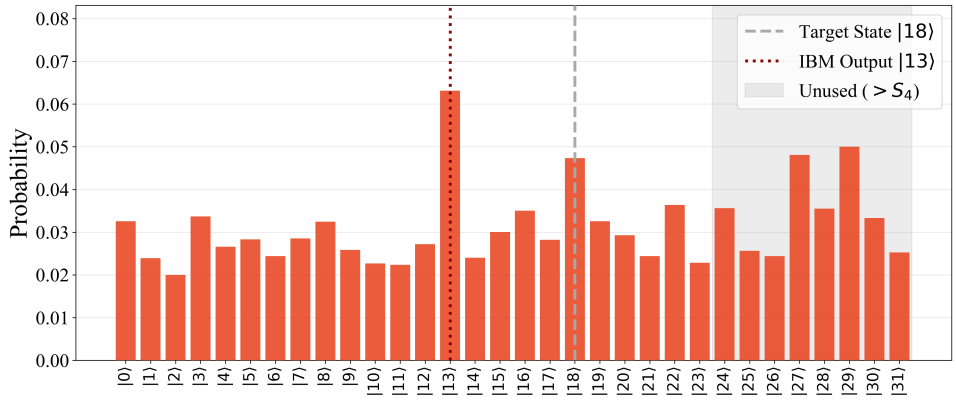
**Note: The S_4 task experienced a single physical training-set failure. However, raw readout analysis confirmed the target remained the distinct second-most probable state (4.7%), retaining a clear probability peak above the 3.1% physical random noise floor.*

time complexity for a sequence length S. Our architecture computes semantic overlap natively via non-Abelian phase interference in superposition. By sequentially composing unitary operators, the quantum state acts as a natively superposed memory buffer. This fundamentally eliminates the need for explicit $\mathcal{O}(S^2)$ attention matrices or multi-qubit fidelity routing gates; instead, the sequence mixing operations scale strictly linearly with the sequence length S and the quantum circuit depth L, achieving an upper-bound time complexity of $\mathcal{O}((S+L)\cdot\log V)$.

Crucially, this logarithmic spatial scaling reveals a profound connection to foundational quantum algorithms. In our modular arithmetic evaluations, the UQT natively learns to decode cyclic, periodic group structures, mathematics fundamentally akin to the phase-estimation logic required for the modular exponentiation step in Shor's algorithm [13]. However, rather than relying on a rigid, general-purpose Inverse Quantum Fourier Transform (IQFT) to extract the



(a) JAX simulation result



(b) Hardware inference result

Figure 4: **Quantum state amplitude distribution for the failed S_4 permutation $P(15) \circ P(17)$. (a)** Ideal probability distribution obtained via noiseless JAX simulation, isolating the target state ($|18\rangle$). **(b)** Raw probability distribution evaluated on the `ibm_marrakesh` physical processor. While unmitigated T_2 decoherence caused a noise-induced bit-flip allowing state $|13\rangle$ to erroneously peak (6.3%), the geometric wave-interference remained physically robust enough to maintain the true target state $|18\rangle$ as the distinct runner-up (4.7%) well above the 3.1% random noise floor.

phase, the parameterized mixing layers organically learn a highly compressed, task-specific basis transformation. This demonstrates that the UQT provides an exceptional parameter efficiency and structural memory advantage for learning the exact discrete algebra that currently forces classical continuous-space models into massive over-parameterization.

DISCUSSION

The ability of a highly compact, 5-qubit physical system to transition between cyclic modular arithmetic and spatial non-Abelian transformations using an identical quantum attention topology suggests that the UQT provides a highly viable substrate for formal mathematical and algebraic reasoning.

Across all empirical evaluations, classical self-attention networks exhibited severe stochastic instability. Because these continuous-space models merely approximate discrete rules, they remain fundamentally vulnerable to the statistical drift that manifests as algorithmic hallucinations. In stark contrast, the UQT physically embodies the target equations, locking into 2π

periodic and $SU(2)$ boundaries. This ability to crystallize into rigid geometric states provides a native mechanism to enforce formal logic deterministically. Rather than relying on increasingly large classical continuous-space networks to approximate discrete logic, these findings suggest that quantum representations can provide structural inductive biases that are well-suited for representing and learning exact formal systems.

REFERENCES

[1] Dziri, N. *et al.* Faith and fate: Limits of transformers on compositionality. *Advances in neural information processing systems* **36**, 70293–70332 (2023).

[2] Trask, A. *et al.* Neural arithmetic logic units. *Advances in neural information processing systems* **31** (2018).

[3] Power, A., Burda, Y., Edwards, H., Babuschkin, I. & Misra, V. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177* (2022).

[4] Liu, Z. *et al.* Towards understanding grokking: An effective theory of representation learning. *Advances in Neural Information Processing Systems* **35**, 34651–34663 (2022).

[5] Gromov, A. Grokking modular arithmetic. *arXiv preprint arXiv:2301.02679* (2023).

[6] Zhang, H. *et al.* A survey of quantum transformers: Architectures, challenges and outlooks. *arXiv preprint arXiv:2504.03192* (2025).

[7] Li, G., Zhao, X. & Wang, X. Quantum self-attention neural networks for text classification. *Science China Information Sciences* **67**, 142501 (2024).

[8] Guo, N. *et al.* Quantum linear algebra is all you need for transformer architectures. *arXiv preprint arXiv:2402.16714* **1** (2024).

[9] Khatri, N., Matos, G., Coopmans, L. & Clark, S. Quixer: A quantum transformer model. *arXiv preprint arXiv:2406.04305* (2024).

[10] Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).

[11] Mitarai, K., Negoro, M., Kitagawa, M. & Fujii, K. Quantum circuit learning. *Physical Review A* **98**, 032309 (2018).

[12] Schuld, M., Bergholm, V., Gogolin, C., Izaac, J. & Killoran, N. Evaluating analytic gradients on quantum hardware. *Physical Review A* **99**, 032331 (2019).

[13] Shor, P. W. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM review* **41**, 303–332 (1999).

METHODS

Mathematical formulation of the UQT. The core function of the UQT relies on encoding classical tokens into geometric quantum states. For an input token x mapped to a parameterized vector $\vec{\theta}_x$, the embedding unitary $\mathcal{E}$ is applied across a dedicated N_{emb}-qubit subspace of the total N-qubit register:

$$\mathcal{E}(\vec{\theta}_x) = \left(\bigotimes_{q=0}^{N_{emb}-1} R_y(\theta_{x,q,3}) R_x(\theta_{x,q,2}) R_y(\theta_{x,q,1}) R_x(\theta_{x,q,0}) \right) \otimes I^{\otimes(N-N_{emb})}, \tag{1}$$

where I represents the identity operation on the remaining un-embedded ancilla qubits. This alternating sequence of orthogonal rotations provides universal single-qubit control, ensuring that each discrete token can be mapped to any arbitrary geometric coordinate on the Bloch sphere. In our empirical evaluations ($N = 5$), we embedded operands into an $N_{emb} = 4$ qubit subspace for modular arithmetic and utilized the full register ($N_{emb} = 5$) for the non-Abelian permutations.

Following the rotational embeddings, the quantum state is processed by digital mixing layers to generate deep structural entanglement. Each layer l applies a parameterized general $SU(2)$ rotation to every qubit, followed by a global cyclic ring of CNOT gates (C_{ring}):

$$U_{\text{mix}}^{(l)} = C_{\text{ring}} \bigotimes_{q=0}^{N-1} U_{\text{Rot}}^{(l)} \left(\alpha_{l,q}, \beta_{l,q}, \gamma_{l,q}\right). \tag{2}$$

In the quantum attention circuit, a full set of tokens $T = (x_1, x_2, \ldots, x_S)$ are embedded consecutively prior to any entanglement, allowing their geometric states to natively superimpose. This state is subsequently processed by L mixing layers to produce the final pre-measurement state:

$$|\psi_{attn}\rangle = \left(\prod_{l=1}^{L} U_{\text{mix}}^{(l)}\right) \left(\prod_{t=1}^{S} \mathcal{E}(\vec{\theta}_{x_t})\right) |0\rangle^{\otimes N}, \tag{3}$$

where the embedding product denotes time-ordered application from right to left.

Optimization and asymmetric regularization. The architecture was optimized using the JAX high-performance array computing framework [14] and the Optax AdamW optimizer. Across all experimental domains, we utilized a constant learning rate of $\eta = 0.005$ alongside moment decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.98$. We applied an asymmetric loss function utilizing an L_2 weight decay ($\lambda_{ent} = 0.01$) exclusively on the entanglement logic gates (W_{ent}), while the phase embedding angles (W_{emb}) were masked from penalization:

$$L_{total} = \mathcal{L}_{\text{NLL}} + \lambda_{ent} ||W_{ent}||_2^2, \tag{4}$$

where W_{ent} encompasses the rotational parameters of the mixing layers, and W_{emb} represents the unpenalized token embedding phases. This architectural constraint permits the token geometry to rotate freely across the Bloch sphere. Simultaneously, the L_2 penalty forces unnecessary entanglement parameters toward zero, functionally reducing arbitrary $SU(2)$ rotations into Identity operations (I). This dynamic natively prunes the circuit during optimization, severely restricting the logical density of the CNOT routing network and preventing over-entanglement.

Dataset generation and architectures. Datasets for modular arithmetic over $\mathbb{Z}_{11}$ and $\mathbb{Z}_{11}^*$, as well as S_4 permutations, were partitioned into 80/20 train/test splits. The UQT utilized a 5-qubit register for these mathematical tasks. The total parameter count for each model is strictly defined by the sum of the phase embedding rotations (W_{emb}) and the parameterized entanglement logic gates (W_{ent}).

To establish a rigorous classical baseline, we trained standard attention-based Transformers using PyTorch [15]. The classical architecture comprised 2 encoder layers, 4 attention heads, an embedding dimension of $d_{\text{model}} = 128$, and a feedforward network dimension of 512. While this represents a relatively small classical network, it requires approximately 4×10^5 trainable parameters. Comparing this to the UQT's sub-700 parameter footprint empirically validates the massive over-parameterization required for classical continuous-space models to approximate discrete formal logic.

Hardware execution. Inference was evaluated on physical superconducting quantum processors accessed via the IBM Qiskit Runtime Service [16]: `ibm_fez` and `ibm_marrakesh` (featuring the 156-qubit Heron r2 architecture). All hardware executions utilized 8,192 measurement shots per evaluation.

Because the physical fidelity of superconducting qubits drifts over time due to continuous environmental decoherence, it is necessary to record the exact baseline hardware metrics at the time of execution to appropriately contextualize the algorithmic noise-resilience of the UQT. The median two-qubit (CZ) gate infidelities across the processors ranged from 2.56×10^{-3} to 2.69×10^{-3}, and median readout assignment errors ranged from 1.05×10^{-2} to 1.52×10^{-2}. Median T_1 relaxation times were bounded between $139.63\,\mu$s and $192.07\,\mu$s, while T_2 dephasing times ranged from $96.32\,\mu$s to $98.48\,\mu$s. Because the UQT does not utilize quantum error correction, the successful, unmitigated hardware classification achieved in our results confirms that the learned geometric phase-interference is robust enough to overcome native environmental decoherence and localized gate degradation.

Why the proposed architecture crystallizes. We now provide a constructive argument showing that the proposed quantum attention architecture can represent modular addition exactly when its embedding operators are chosen to respect the symmetry of the task. While the following discussion is focused on modular arithmetic, the extension to non-abelian algebra is straightforward. Consider the function

$$f(n, m) = n + m \pmod{p}, \tag{5}$$

where $n, m \in \mathbb{Z}_p$ and $p \geq 2$ is the modulus. The key observation is that modular addition is the group operation of the finite cyclic group $\mathbb{Z}_p$ [5]. Therefore, the natural basis for representing this task is given by the characters of $\mathbb{Z}_p$, namely

$$\chi_k(n) = e^{2\pi i k n/p}, \qquad k = 0, \ldots, p-1. \tag{6}$$

These functions diagonalize the translation structure of the problem and satisfy

$$\chi_k(n)\chi_k(m) = \chi_k(n+m), \tag{7}$$

where addition is understood modulo p. This multiplicative identity is the central reason the architecture can implement modular addition efficiently.

To analyze what happens at the embedding stage, let $\{|k\rangle\}_{k=0}^{p-1}$ denote a computational basis over a p-dimensional latent register. Define the token embedding operator by

$$U_{\text{emb}}(n)|k\rangle = \chi_k(n)|k\rangle = e^{2\pi i k n/p}|k\rangle. \tag{8}$$

Thus each token n is encoded as a phase shift across spectral modes. Unlike unconstrained Euclidean embeddings, this representation is periodic and exactly compatible with modular arithmetic. By initializing the latent register in the uniform superposition, i.e.,

$$|\psi_0\rangle = \frac{1}{\sqrt{p}} \sum_{k=0}^{p-1} |k\rangle, \tag{9}$$

and applying the embedding for token n, we have

$$|\psi(n)\rangle = U_{\text{emb}}(n)|\psi_0\rangle = \frac{1}{\sqrt{p}} \sum_{k=0}^{p-1} e^{2\pi i k n/p}|k\rangle. \tag{10}$$

At the composition step, applying two token embeddings sequentially yields

$$U_{\text{emb}}(m)U_{\text{emb}}(n)|k\rangle = e^{2\pi i k m/p} e^{2\pi i k n/p}|k\rangle \tag{11}$$

$$= e^{2\pi i k (n+m)/p}|k\rangle. \tag{12}$$

Therefore,

$$U_{\text{emb}}(m)U_{\text{emb}}(n)|\psi_0\rangle = \frac{1}{\sqrt{p}} \sum_{k=0}^{p-1} e^{2\pi i k (n+m)/p}|k\rangle. \tag{13}$$

The modular sum is thus accumulated directly in phase space. No learned nonlinear arithmetic rule is required; the group structure of the task is implemented by construction.

Finally, at the readout stage, to decode the symbolic output, we can apply the IQFT ($F_p^\dagger$):

$$F_p^\dagger |k\rangle = \frac{1}{\sqrt{p}} \sum_{q=0}^{p-1} e^{-2\pi i k q/p} |q\rangle. \tag{14}$$

Then,

$$F_p^\dagger \left(\frac{1}{\sqrt{p}} \sum_{k=0}^{p-1} e^{2\pi i k(n+m)/p} |k\rangle \right) = \sum_{q=0}^{p-1} \left[\frac{1}{p} \sum_{k=0}^{p-1} e^{2\pi i k(n+m-q)/p} \right] |q\rangle. \tag{15}$$

Using orthogonality of characters,

$$\sum_{k=0}^{p-1} e^{2\pi i k s/p} = p\, \delta_{s,0 \,(\mathrm{mod}\, p)}, \tag{16}$$

we obtain

$$F_p^\dagger U_{\mathrm{emb}}(m) U_{\mathrm{emb}}(n) |\psi_0\rangle = |n + m \pmod p\rangle. \tag{17}$$

Accordingly, measurement returns the correct modular sum with probability one. While this analytical construction relies on an explicitly programmed IQFT and exact character phase shifts, it serves as a foundational existence proof: the exact geometry of modular arithmetic is natively resolvable within a multi-qubit Hilbert space. Classical Euclidean models lack this periodic, wave-interference inductive bias, forcing them to approximate these dynamics through massive over-parameterization.

Note that in our proposed UQT, we do not hardcode the IQFT nor the exact $\mathbb{Z}_p$ group characters. Instead, the architecture utilizes parameterized $SU(2)$ rotations (R_x and R_y gates) and unconstrained gradient optimization to organically discover a task-specific basis transformation. The phenomenon of crystallization we observe empirically suggests that the UQT learns an isomorphic, highly compressed representation of the phase-accumulation and decoding process described above, ultimately achieving the same deterministic, probability-one generalization. By leveraging the continuous, periodic nature of quantum phase space, the UQT natively converges to the target algebraic structure without the stochastic approximation errors inherent to classical continuous-space networks.

It is also important to note the key difference between the aforementioned discussion and the argument presented by Gromov [5]. In their work, they utilized real-valued cosine features as an approximate Fourier basis (unlike the proposed method that employs the full complex characters, $e^{2\pi i k n/p}$). The cosine formulation captures only the real projection of the harmonic structure, effectively discarding the imaginary sine component that carries phase orientation and completes the representation. As a result, the MLP must reconstruct modular relations indirectly through nonlinear interactions and training dynamics, whereas the proposed architecture preserves the full phase information from the outset, allowing modular composition to arise naturally through direct multiplication of complex phases rather than through delayed emergence of partial real-valued harmonics.

METHODS REFERENCES

[14] Bradbury, J. *et al.* JAX: composable transformations of Python+NumPy programs (2018). URL http://github.com/jax-ml/jax.

[15] Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019).

[16] Javadi-Abhari, A. *et al.* Quantum computing with qiskit. *arXiv preprint arXiv:2405.08810* (2024).

AUTHOR CONTRIBUTIONS

S.C. conceived the Universal Quantum Transformer architecture and designed its quantum circuit topologies, implemented the JAX and PyTorch codebases, executed the inference evaluations on physical IBM Quantum hardware, and wrote the original manuscript. A.T. supervised the research, collaborated in advancing and refining the initial version of the model architecture, contributed to the theoretical analysis, validated the algorithms and code, and critically revised the writing.

FUNDING DECLARATION

The authors declare that they have no funding sources to disclose for this work.

