Quantum-Inspired Complex Transformers: Resolving the Fundamental Algebraic Ambiguity for Enhanced Neural Representations

NeurIPS 2025 Workshop (NEGEL)

Bhargav Patel

Cambrian College
Laurentian University
(Graduated Masters Student)

NEURAL INFORMATION PROCESSING SYSTEMS | 2025

LaurentianUniversity UniversitéLaurentienne

Complexnumber Ambiguity

- Complex numbers (i) form the foundation of Quantum Mechanics.

But what if the **complex number itself** could exist in a **state of superposition**?

Thought Experiment: Origin of the Imaginary Unit

Consider an equation with two real-structured solutions:

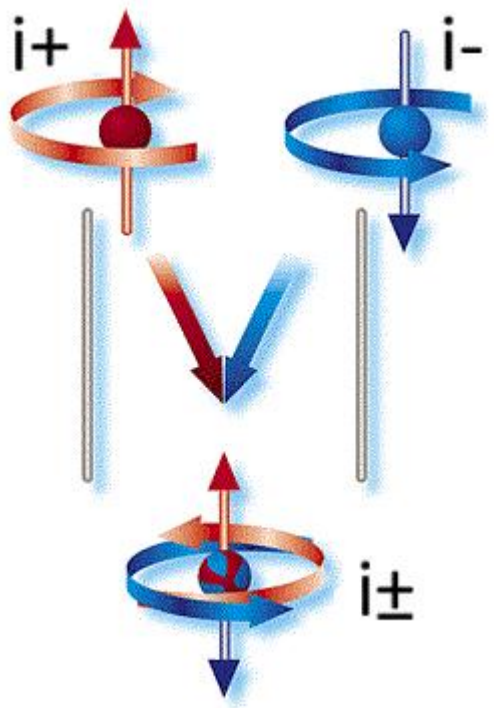
$$Y = 1 \pm \sqrt{2}$$

Now, let's multiply these two conjugate-like terms:

$$Y_+ \cdot Y_- = (1+\sqrt{2})(1-\sqrt{2}) = 1 - 2 = -1$$

Thus,

$$Y^2 = -1 \implies Y = \pm i$$



- This suggests that the imaginary unit (i) can be interpreted as emerging from the interaction (or product) of two opposing real-valued roots.
- Hence, instead of considering just i, we might view **Y = ±i as the more** fundamental representation - a dual or superposed origin of the imaginary dimension.

Complex 1D Conjugate to 2D Superposition

Consider the general 1D conjugate family:

$$Y = \alpha \pm \beta, \quad \alpha, \beta \in \mathbb{R}$$

- Infinite solutions exist for any α and β
- The product of the conjugates is:

$$Y_+ \cdot Y_- = (\alpha + \beta)(\alpha - \beta) = \alpha^2 - \beta^2 = -1$$

- However, these solutions **do not satisfy the properties of the imaginary unit**:

$$+i \cdot -i = 1, \quad +i^2 = -1, \quad -i^2 = -1$$

This indicates that ±i cannot exist as a single 1D number.

In 2D, ±i can be represented as matrices:

$$J_+ = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad J_- = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

- Squaring reproduces the imaginary unit:

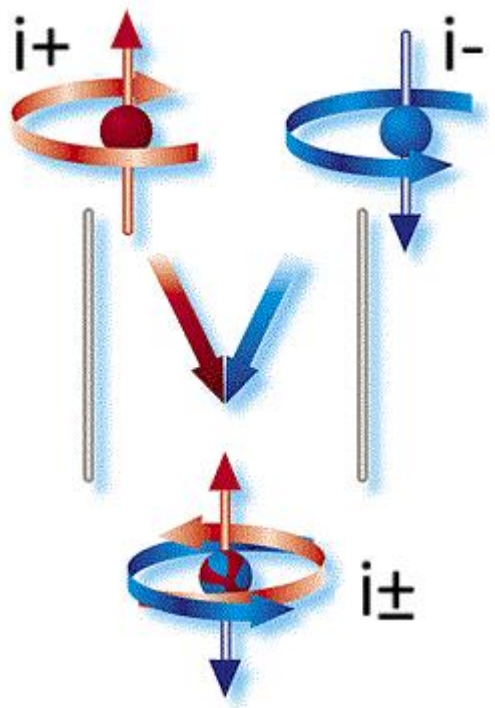
$$J_+ \cdot J_+ = -I \sim +i^2, \quad J_- \cdot J_- = -I \sim -i^2$$

- Multiplying the two matrices mimics the 1D conjugate product:

$$J_+ \cdot J_- = I \sim (+i) \cdot (-i)$$

- Observing J_+ and J_-, the -1 appears to **fluctuate between two axes**

This naturally raises the question: what if J_+ and J_- exist in a **quantum superposition**?



Complex 1D Conjugate to 2D Superposition

Define the continuous 2D superposition:

$$J(t) = \cos t\, J_+ + \sin t\, J_-$$

Expansion using matrix form:

$$J_+ = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad J_- = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

$$J(t) = \cos t \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} + \sin t \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & \cos t - \sin t \\ \sin t - \cos t & 0 \end{bmatrix}$$

Generalization:

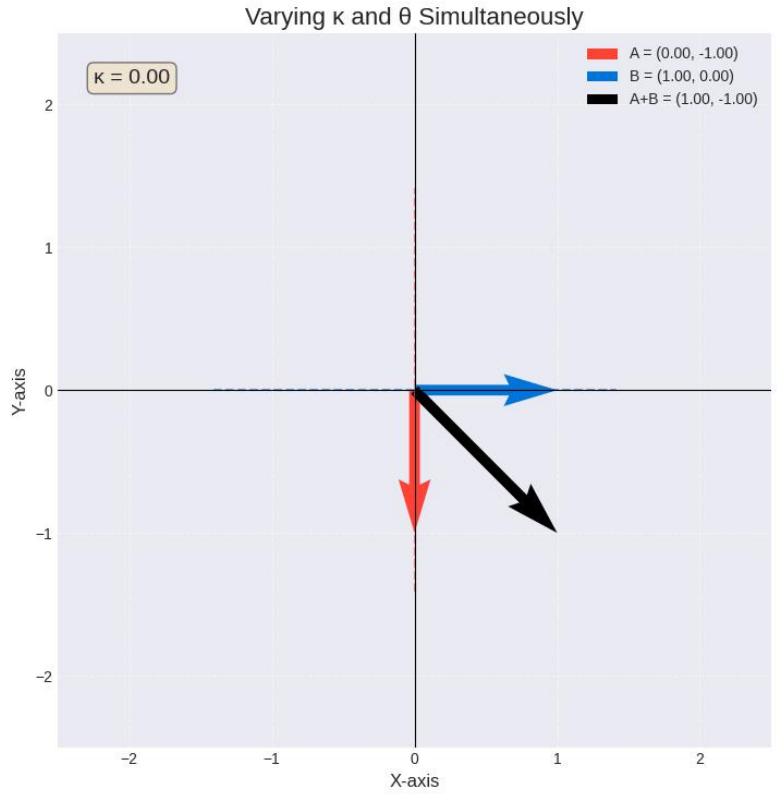
$$\begin{bmatrix} a & b(\cos t - \sin t) \\ b(\sin t - \cos t) & a \end{bmatrix}, \quad a, b \in \mathbb{R}$$

- ±i in 2D can be **continuously rotated and superposed**
- Coefficients a, b allow representation of **generalized 2D operators**

Quantum Complex Number:

$$\boxed{Z = a + b\, J(t)}$$

- This represents a **time-dependent, quantum-like complex number** in 2D



Transformers and Quantum Complex Numbers

QIC Linear Layers

The fundamental building block extends matrix multiplication to QIC algebra. For input $x = x_a + x_b J$ and weights $W = W_a + W_b J$:

$$y = Wx + b \tag{13}$$

$$= [W_a x_a + W_b x_b(-1 + \sin(2\theta)) + b_a] + [W_a x_b + W_b x_a + b_b]J \tag{14}$$

Implementation maintains separate real and imaginary components, with interactions governed by the learnable θ.

QIC Attention Mechanism

For QIC attention with queries Q, keys K, and values V, we compute attention scores as $S = QK^T = S_a + S_b J$, apply softmax to obtain attention weights $\alpha_{ij} = \frac{\exp(|S_{ij}|/\sqrt{d_k})}{\sum_k \exp(|S_{ik}|/\sqrt{d_k})}$, and aggregate values as $\text{Attention}(Q, K, V) = \alpha V_a + \alpha V_b J$.

Multi-head attention uses head-specific phase parameters θ_h, allowing different heads to operate in different algebraic regimes:

$$\text{head}_h = \text{Attention}_{\theta_h}(QW_h^Q, KW_h^K, VW_h^V) \tag{15}$$

$$\text{QIC-LayerNorm}(z) = \gamma \frac{z - \mu}{\|\sigma\|_2} \tag{16}$$

where μ and σ are computed over the magnitudes $|z_i|$ across the normalized dimension.

For activation functions, we adopt magnitude-based nonlinearities that preserve the QIC structure, inspired by the success of gated linear units [23]:

$$\text{QIC-ReLU}(z) = \text{ReLU}(|z|) \cdot \frac{z}{|z|} \tag{17}$$

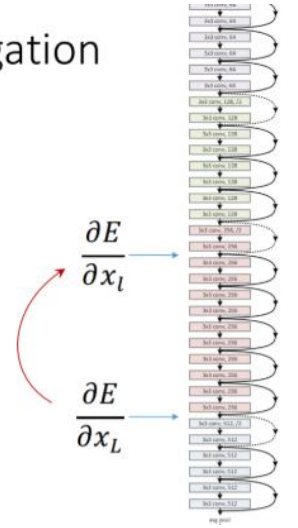
Residual Like Connection: while computing gradient

$$\frac{\partial \mathcal{L}}{\partial \theta} = 2\cos(2\theta) \sum_{i,j} \frac{\partial \mathcal{L}}{\partial y_{a,ij}} W_{b,ij} x_{b,ij}$$

Very smooth backward propagation

$$\frac{\partial E}{\partial x_l} = \frac{\partial E}{\partial x_L}\left(1 + \frac{\partial}{\partial x_l}\sum_{i=1}^{L-1} F(x_i)\right)$$

- Any $\frac{\partial E}{\partial x_L}$ is directly back-prop to any $\frac{\partial E}{\partial x_l}$, plus residual.
- Any $\frac{\partial E}{\partial x_l}$ is additive; unlikely to vanish
  - in contrast to multiplicative: $\frac{\partial E}{\partial x_l} = \prod_{i=l}^{L-1} W_i \frac{\partial E}{\partial x_L}$



Residual Networks (ResNets, He et al., 2015)

Results

Model Parameters: Standard = 1,466,370 | QIC = 774,407 (−47.2%)

Dataset	Metric	Standard	QIC	Improvement
IMDB Sentiment	Accuracy	100.0%	100.0%	—
	Training Time/Epoch	115.1s	102.7s	−10.8%
AG News	Accuracy	73.3%	78.0%	+4.7%
	Final Loss	0.4056	0.4066	+0.2%
Overall	Avg. Accuracy	86.7%	89.0%	+2.3%

Highlights:

- QIC achieves 47.2% fewer parameters (774K vs 1.47M).
- 10.8% faster training per epoch.
- 4.7% higher accuracy on multi-class task.
- Demonstrates quantum-inspired efficiency without loss of performance.

Vaswani et al., "Attention Is All You Need," NeurIPS 2017.

Abliation Studies

Learned vs. Fixed θ: Fixing $\theta = \pi/4$ reduces accuracy by 2.8%, demonstrating that learning the algebraic unit is crucial. When $\theta = 0$ (equivalent to standard complex numbers with fixed i), accuracy drops by 3.2%, confirming that the learnable superposition provides genuine benefits beyond fixed complex arithmetic.

Table 2: Ablation study results on AG News dataset

Configuration	Accuracy	Parameters	Analysis
Full QIC Transformer	78.0%	774,407	Full model
Fixed $\theta = \pi/4$	75.2%	774,396	-2.8% accuracy
Fixed $\theta = 0$ (standard complex)	74.8%	774,396	-3.2% accuracy
Global θ (not per-head)	76.4%	774,401	-1.6% accuracy
Parameter-matched real baseline	73.1%	774,400	-4.9% accuracy
Standard Transformer	73.3%	1,466,370	2× parameters

Scope of θ: Using a single global θ instead of per-head parameters reduces accuracy by 1.6%, validating that different attention heads benefit from operating in different algebraic regimes.

Initialization Sensitivity: We tested three initializations: $\theta = 0$, $\theta = \pi/4$, and random $\theta \sim \mathcal{U}(0, \pi/2)$. All converged to similar final accuracy ($\pm 0.3\%$), with final θ values clustering around 0.75-0.85 regardless of initialization, suggesting a learnable optimum.

Comparisons & Parameters Sensitivity

Model	Accuracy	Parameters
Standard Real Transformer	73.3%	1,466,370
Complex Transformer (fixed i)	74.8%	774,396
Complex Transformer (i with phase gates)	75.6%	812,450
Quaternion Transformer	75.1%	806,200
QIC Transformer (ours)	**78.0%**	**774,407**

B.3 Hyperparameter Sensitivity

We tested sensitivity to key hyperparameters:

Learning Rate: Tested $\{10^{-4}, 5\times10^{-4}, 10^{-3}, 5\times10^{-3}\}$. QIC performance stable across range, with optimum at 10^{-3} (same as standard). QIC shows slightly wider stable range.

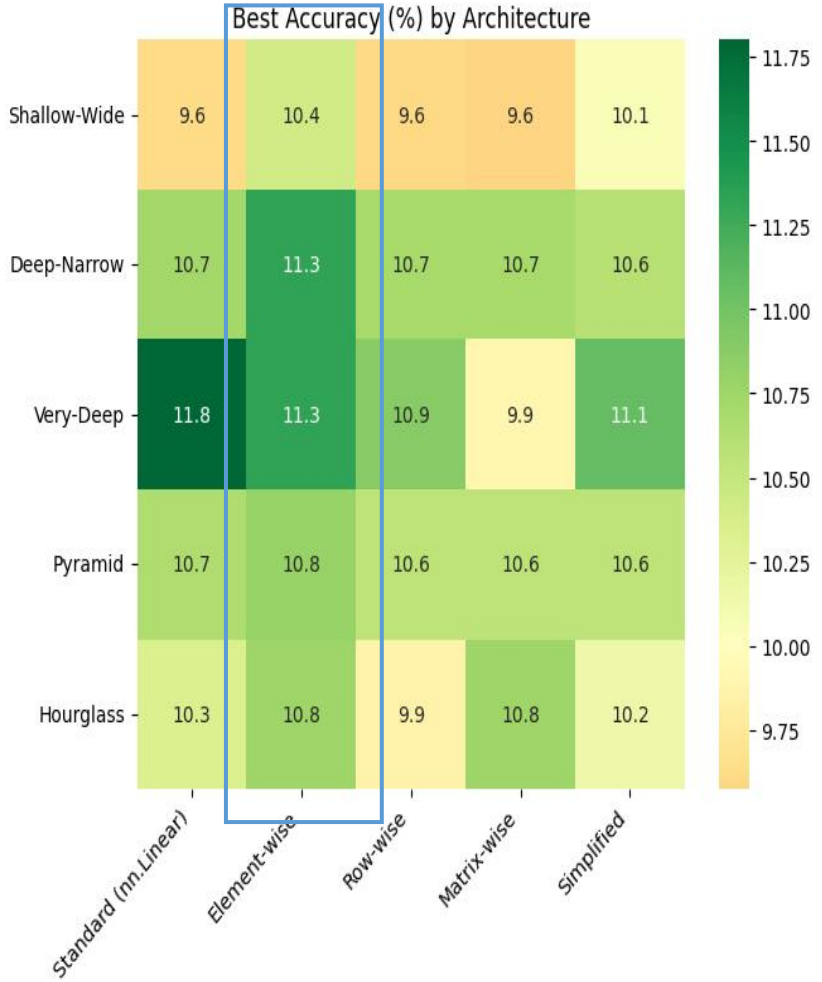
Batch Size: Tested $\{16, 32, 64, 128\}$. Performance similar across range. Memory advantage of QIC more pronounced at larger batch sizes.

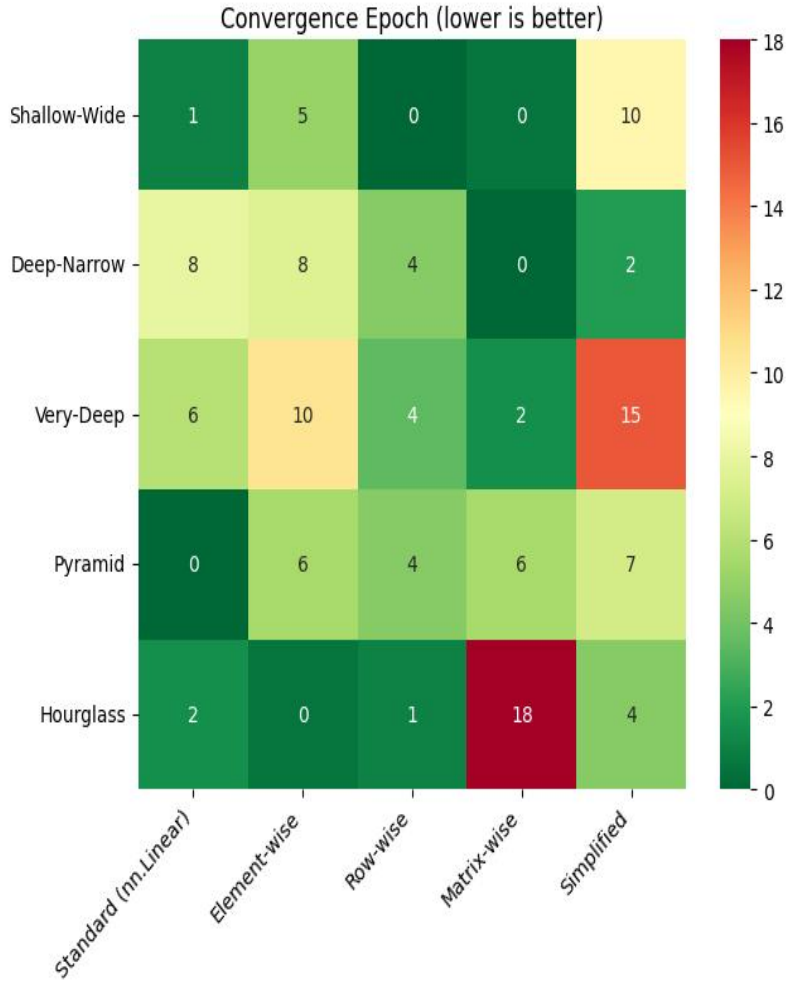
Phase Parameter Initialization: Tested $\theta_0 \in \{0, \pi/6, \pi/4, \pi/3, \text{random}\}$. All converged to similar final performance ($\pm0.3\%$) and similar final θ values (0.75-0.85), indicating robust learning dynamics.

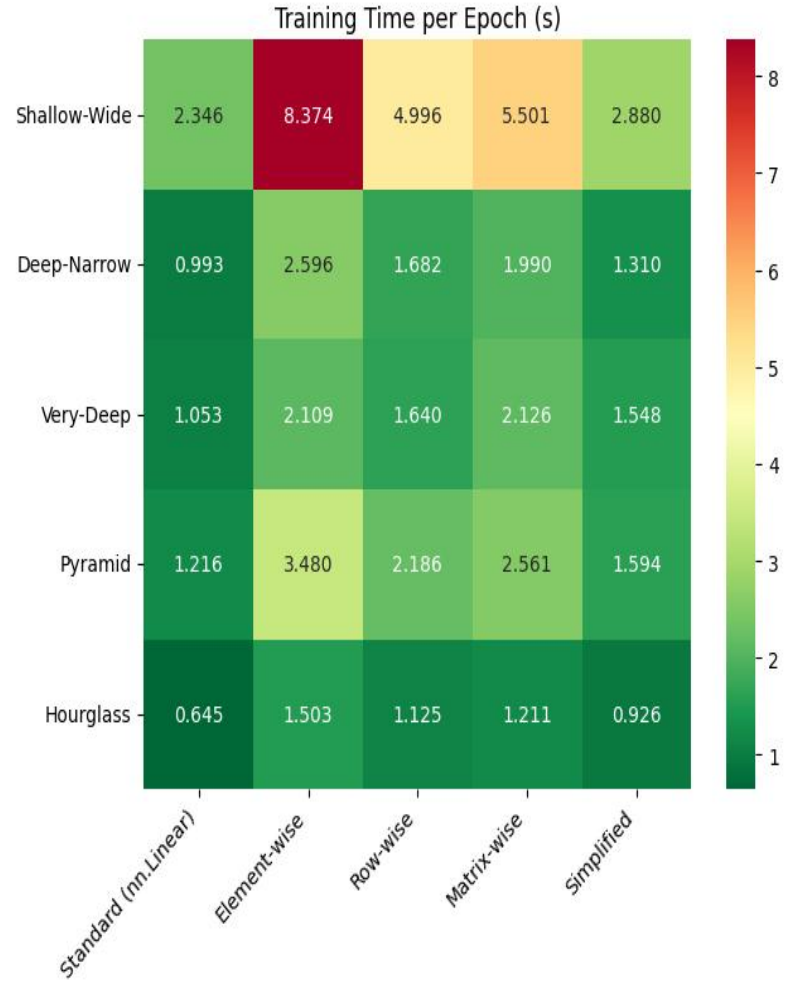
> *Baselines reproduced from Vaswani et al. (2017); Trabelsi et al. (2018); Chi et al. (2020); Parcollet et al. (2019, 2021).*

Future Work / Work In Progress

Discovered a more **stable layer, analogous to a linear layer** (W^TX), in neural networks with quantum complex numbers $J(\theta)$; tested on CIFAR-10 and extending toward more powerful transformers.







Future Work / Work In Progress

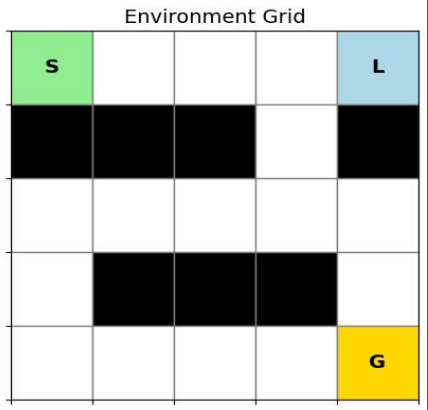
We experimented with variants of a stable layer analogous to a linear layer (W^TX), drawing research direction from this quantum-complex phenomenon, validated on CIFAR-100, and extended toward more powerful neural net architecture.

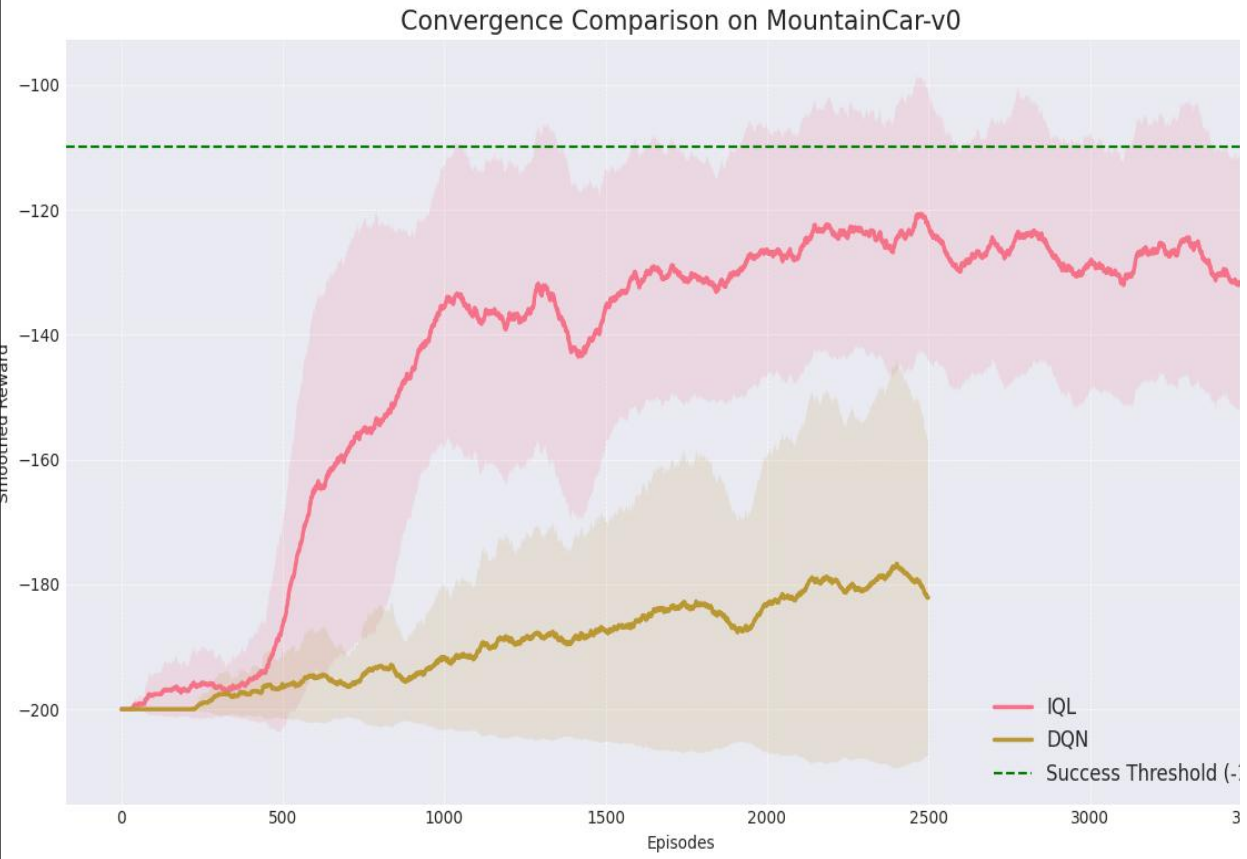
Dataset	Architecture	Standard MLP	New -Linear	Δ	Relative
CIFAR-100	Small [256]	27.59 ± 0.21	**29.77 ± 0.34**	+2.19	+7.9%
CIFAR-100	Medium [512-256]	26.95 ± 0.26	**34.08 ± 0.14**	+7.14	+26.5%
CIFAR-100	Large [1024-512-256]	25.02 ± 0.25	**33.51 ± 0.16**	+8.49	+33.9%
SVHN	Small [256]	76.97 ± 0.12	**79.59 ± 0.16**	+2.62	+3.4%
SVHN	Medium [512-256]	**80.49 ± 0.26**	80.29 ± 0.17	−0.20	−0.2%
SVHN	Large [1024-512-256]	82.54 ± 0.11	**86.10 ± 0.25**	+3.56	+4.3%

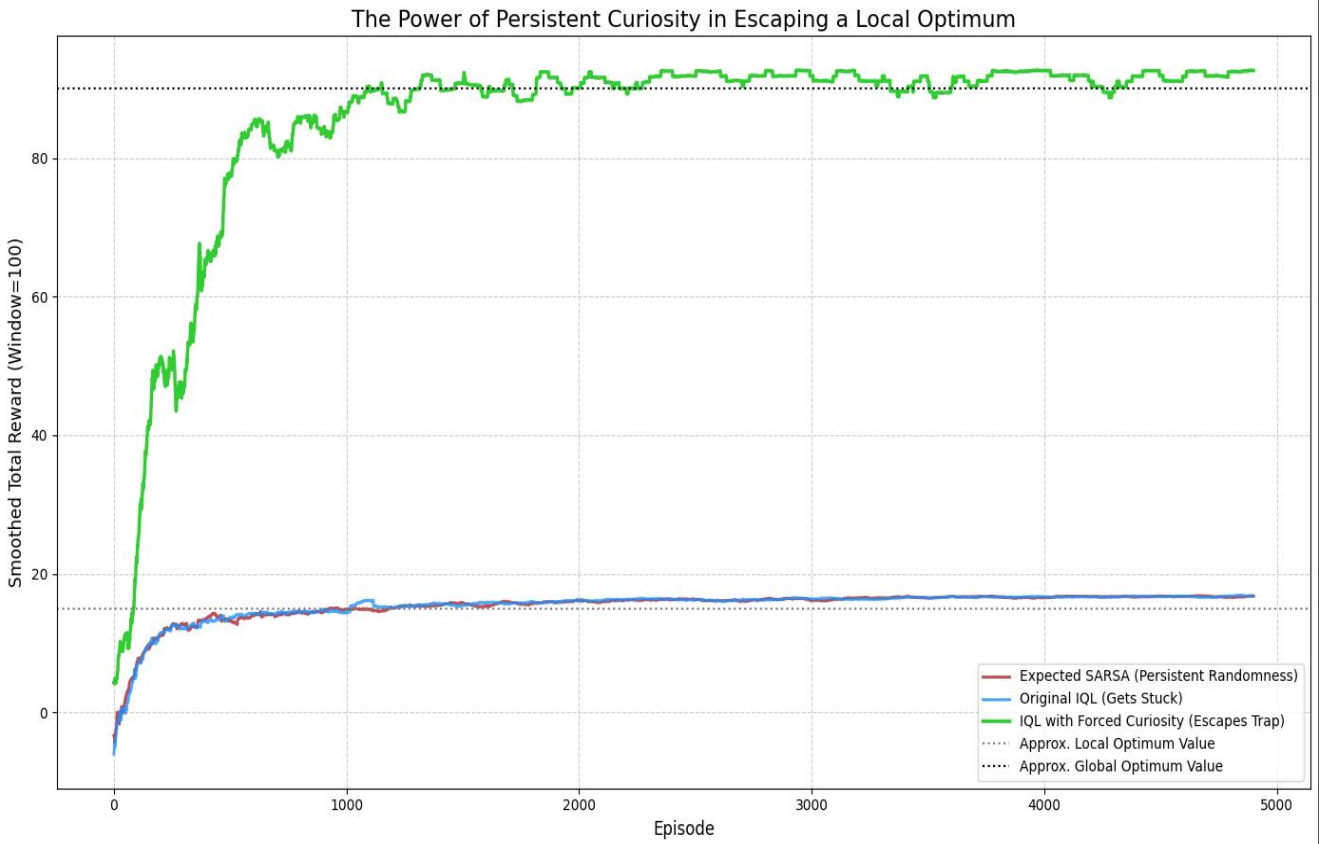
Future Work / Work In Progress

Discovered an off-policy Auto-Explore & Exploit algorithm (**eliminating the epsilon-greedy method**) leveraging the Quantum Complex Number J(θ) phenomenon; currently taking small steps toward shaping the future of RL in LLMs.

IQL is Middle Ground between UCB and Q-Learning:
No future step lookups (as in UCB) &
No random action selection (as in Q-Learning)







Thank You!

Exploring challenging AI opportunities in Canada/US

b.patel.physics@gmail.com