Results

This document contains some results obtained by analyzing the data along with calculation and conclusions to support them.

1. Assessment of Quantum Advantage in Hybrid Quantum-Classical CNNs

- MNIST

  Clean Accuracy:
  - Classical CNN Accuracy: 0.989
  - Best HQCNN Accuracy: 0.992 (cnot, circular)
  - Delta: +0.003 → Small but positive quantum advantage

  FGSM:
  - Classical: 0.986
  - Best HQCNN: 0.986 (tied) — No gain

  PGD:
  - Classical: 0.987
  - Best HQCNN: 0.981–0.975 → Slight drop in robustness

  Conclusion (MNIST):
  Quantum advantage exists in clean accuracy, but not in adversarial robustness. The dataset's simplicity likely leads to saturation in classical model performance, leaving little room for improvement.

- Fashion-MNIST

  Clean Accuracy:
  - Classical CNN: 0.915
  - Best HQCNN: 0.922 (cz, linear)
  - Delta: +0.007 → Meaningful advantage

  FGSM:
  - Classical: 0.748
  - Best HQCNN: 0.801 (cz, circular)
  - Delta: +0.053 (+5.3%)

  PGD:
  - Classical: 0.709
  - Best HQCNN: 0.75 (cz, circular)
  - Delta: +0.041 (+4.1%)

  Conclusion (FMNIST):
  Clear and consistent quantum advantage in both clean and adversarial scenarios. This suggests HQCNNs extract more meaningful structure from mid-complexity data and generalize better even under attack.

- CIFAR-10

  Clean Accuracy:
  - Classical CNN: 0.713
  - Best HQCNN: 0.702 (cz, circular)
  - Delta: −0.011 → No quantum advantage in clean accuracy

FGSM:

- Classical: 0.173
- Best HQCNN: 0.258 (cnot, full)
- Delta: +0.085 (+8.5%)

PGD:

- Classical: 0.122
- Best HQCNN: 0.191 (cnot, full)
- Delta: +0.069 (+6.9%)

Conclusion (CIFAR-10):

Quantum models show a large advantage in robustness to adversarial attacks, despite lower clean accuracy. This suggests quantum-enhanced architectures may prioritize more robust features, even at the cost of baseline performance in more complex datasets.

Observations Matrix for Quantum Advantage:

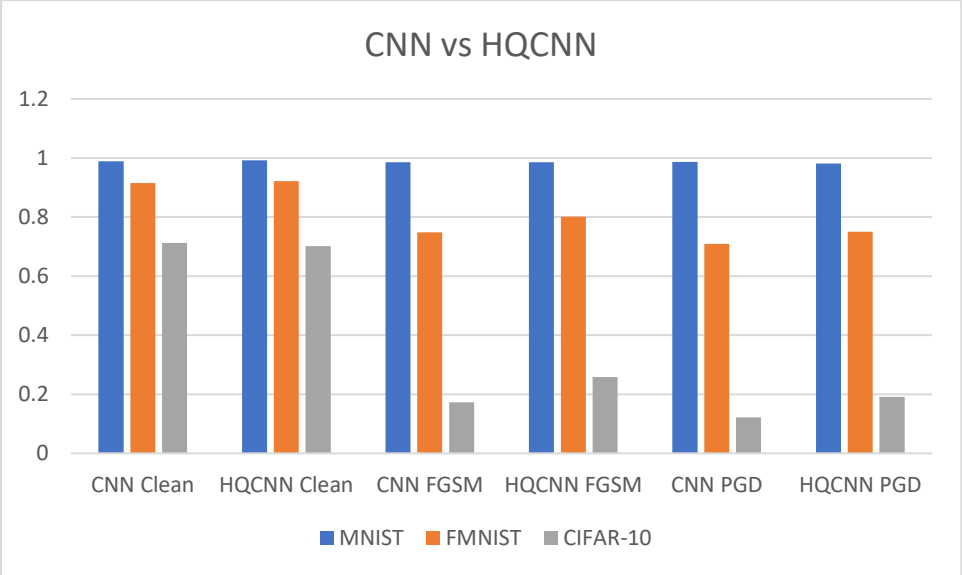
	Clean	FGSM	PGD
MNIST	✅ Small	❌ None	❌ None
FMNIST	✅ Clear	✅ Large	✅ Large
CIFAR-10	❌ None	✅ Strong	✅ Strong

Overall Conclusion:

Quantum advantage is data and task-dependent. HQCNNs offer clear benefits:

- On mid-complexity data (FMNIST): in both accuracy and robustness
- On high-complexity data (CIFAR-10): in robustness only
- On low-complexity data (MNIST): little advantage due to ceiling effect

These findings suggest that quantum models are most useful when classical models struggle — either due to data complexity or attack vulnerability.

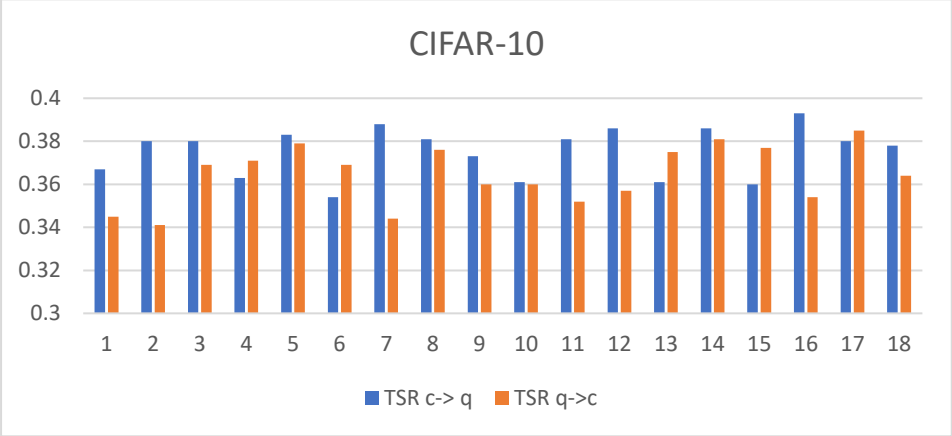


2. Statistical Evidence Favors Classical → Quantum Transfer Attacks

The strength of transfer attacks is measured by the Transfer Success Rate (TSR).

$$TSR_{\{A \to B\}} = \frac{\text{(Number of adversarial examples that fool B)}}{\text{(Total adversarial examples from A)}}$$

- CIFAR-10 dataset:



t-Test: Paired Two Sample for Means

	TSR c-> q	TSR q->c
Mean	0.375277778	0.364388889
Variance	0.000130801	0.000182369
Observations	18	18
Pearson Correlation	-0.197647418	
Hypothesized Mean Difference	0	
df	17	
t Stat	2.388113876	
P(T<=t) one-tail	0.014405192	
t Critical one-tail	1.739606726	
P(T<=t) two-tail	0.028810385	
t Critical two-tail	2.109815578	

Conclusion (One-Tailed Test)

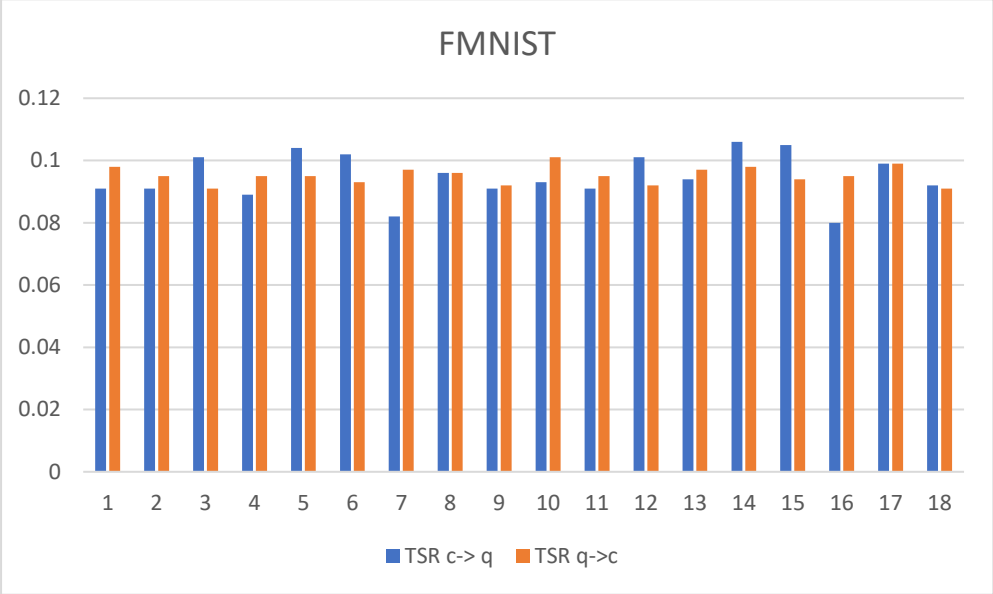
- The **null hypothesis (H_0)** is that the mean of TSR c→q is **less than or equal to** the mean of TSR q→c.
- The **alternative hypothesis (H_1)** is that TSR c→q is **greater than** TSR q→c.

Since:

- **p-value (0.0144) < 0.05**, and
- **t Stat (2.3881) > t Critical one-tail (1.7396)**,

Hence, **we reject the null hypothesis** at the **5% significance level**. Thus, there is **statistically significant evidence** to conclude that **TSR c→q is greater than TSR q→c** for CIFAR-10.

- FMNIST dataset:



t-Test: Paired Two Sample for Means

	TSR c-> q	TSR q->c
Mean	0.094888889	0.095222222
Variance	5.57516E-05	7.83007E-06
Observations	18	18
Pearson Correlation	-0.122626164	
Hypothesized Mean Difference	0	
df	17	
t Stat	-0.170615325	
P(T<=t) one-tail	0.433270197	
t Critical one-tail	1.739606726	
P(T<=t) two-tail	0.866540395	
t Critical two-tail	2.109815578	

Conclusion (One-Tailed Test)

- **The null hypothesis (H_0)** is that the mean of **TSR c→q** is **less than or equal to** the mean of **TSR q→c**.
- **The alternative hypothesis (H_1)** is that **TSR c→q is greater than TSR q→c**.

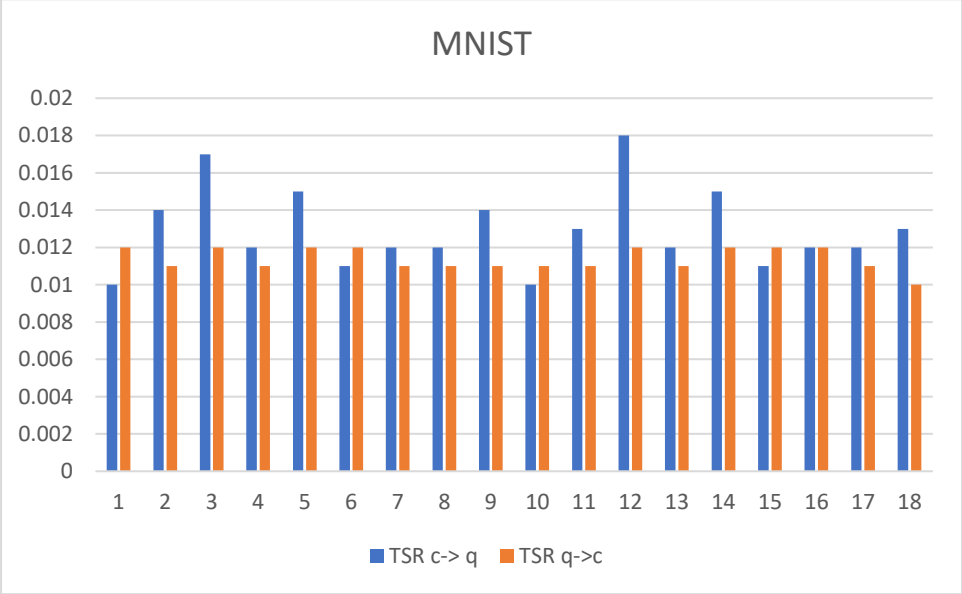
Since:

- **p-value (0.4333) > 0.05**, and
- **t Stat (-0.1706) < t Critical one-tail (1.7396)**,

Hence**, we fail to reject the null hypothesis** at the 5% significance level.

There is **no statistically significant evidence** to conclude that **TSR c→q is greater than TSR q→c**.
In fact, the negative t-statistic suggests the **opposite direction**, though not significantly — **TSR q→c may be slightly higher**, but the difference is not statistically meaningful.

- MNIST dataset:



t-Test: Paired Two Sample for Means

	TSR c-> q	TSR q->c
Mean	0.012944444	0.011388889
Variance	4.87908E-06	3.69281E-07
Observations	18	18
Pearson Correlation	0.236157971	
Hypothesized Mean Difference	0	
df	17	
t Stat	3.072310773	
P(T<=t) one-tail	0.00345094	
t Critical one-tail	1.739606726	
P(T<=t) two-tail	0.00690188	
t Critical two-tail	2.109815578	

Conclusion (One-Tailed Test)

- **The null hypothesis (H_o)** is that the mean of **TSR c→q** is **less than or equal to** the mean of **TSR q→c**.
- **The alternative hypothesis (H_1)** is that **TSR c→q is greater than TSR q→c**.
  Since:
- **p-value (0.0035) < 0.05**, and
- **t Stat (3.0723) > t Critical one-tail (1.7396)**,

Hence, **we reject the null hypothesis** at the 5% significance level. Thus, there is **statistically significant evidence** to conclude that **TSR c→q is greater than TSR q→c**, for MNIST.

In 2 out of 3 datasets, **TSR c→q** statistically significantly greater than **TSR q→c**. However, in the second dataset, no significant difference was observed. Therefore, we **cannot confidently generalize** that **TSR c→q** is greater than **TSR q→c** across all datasets without further analysis.

- Aggregate Results:

A final t-test can be performed for the aggregate data across all three datasets. This gives the following results:

t-Test: Paired Two Sample for Means

	TSR c-> q	TSR q->c
Mean	0.161037037	0.157
Variance	0.024584225	0.023165396
Observations	54	54
Pearson Correlation	0.996920293	
Hypothesized Mean Difference	0	
df	53	
t Stat	2.288280449	
P(T<=t) one-tail	0.013068605	
t Critical one-tail	1.674116237	
P(T<=t) two-tail	0.02613721	
t Critical two-tail	2.005745995	

Conclusion (One-Tailed Test)

- **The null hypothesis (H_o)** is that the mean of **TSR c→q** is **less than or equal to** the mean of **TSR q→c**.
- **The alternative hypothesis (H_1)** is that **TSR c→q is greater than TSR q→c**.

Since:

- The **one-tailed p-value = 0.01307 < 0.05**
- The **t Stat = 2.28828 > 0**
- The **t Stat > t Critical (1.674)**

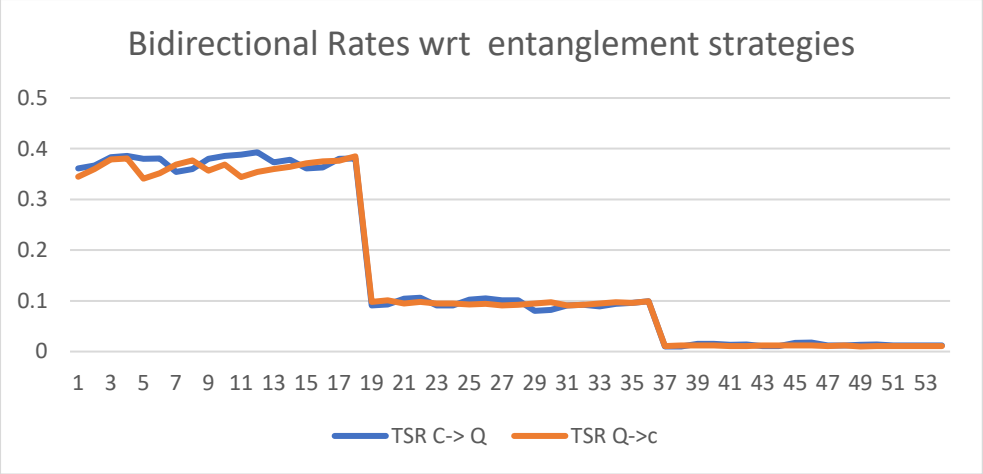
Hence, **we reject the null hypothesis** and conclude that **TSR c→q is statistically significantly greater than TSR q→c** at the 5% significance level, when considering the combined data from all three datasets.

Based on the aggregate paired t-test across all 54 paired observations from the three datasets, there is statistically significant evidence (**p = 0.013**) that **TSR c→q is greater than TSR q→c** overall.

Overall Conclusion:

Transfer attacks from Classical CNNS to Hybrid Quantum CNNs appear to be a greater risk than transfer attacks from Hybrid Quantum CNNs to Classical CNNs. This asymmetry in transferability has heavy security implications.

In practical adversarial settings, attackers may not have access to quantum models but can easily target classical ones. These results imply that such attacks can still **compromise quantum systems indirectly**.



---

3. No Statistical Evidence for correlation between Transferability and Robustness

To find a correlation between these two variables we can perform Spearman's correlation test.

- **Null Hypothesis (H_0):** There is **no monotonic relationship** between accuracy (acc) and incoming TSR, i.e. the Spearman correlation coefficient $\rho = 0$.
- **Alternative Hypothesis (H_1):** There **is a monotonic relationship** between accuracy (acc) and incoming TSR, i.e. the Spearman correlation coefficient $\rho \neq 0$.

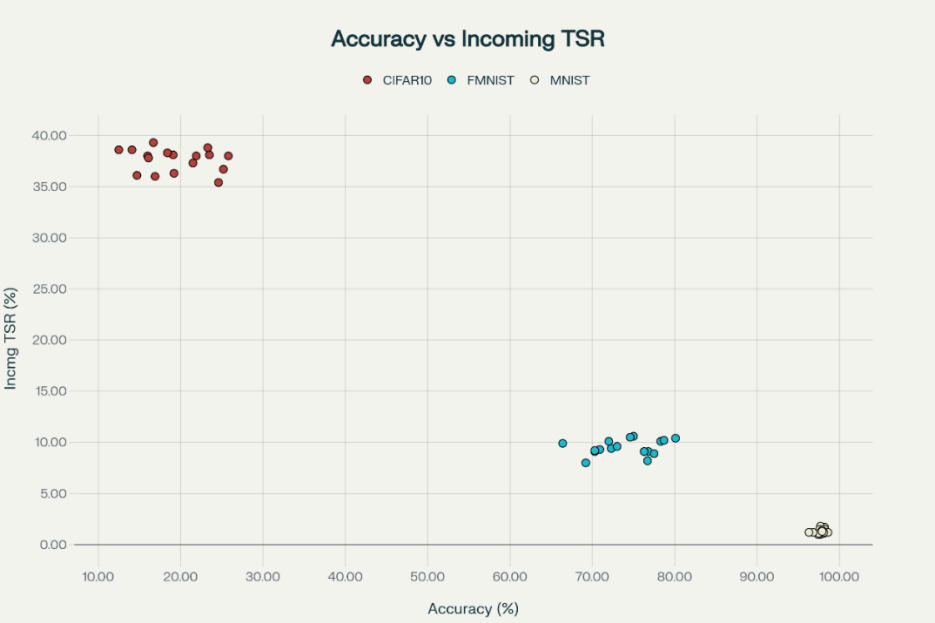
The Aggregate Spearman's correlation test between accuracy (acc) and incoming TSR gives the following results:

- Spearman correlation coefficient: -0.879
- p-value: 2.59×10^{-18}

Interpretation:

- The strong **negative correlation** ($\rho \approx -0.88$) indicates that as incoming TSR increases, accuracy tends to decrease.
- The **extremely low p-value** confirms that this correlation is statistically significant.

However, a scatter plot of all the data reveals 3 clusters for each of the 3 datasets.



- **Caveat:**
  This strong negative correlation is driven by differences between datasets, not within each dataset. Within each dataset, the correlation is weak or absent. Performing Correlation tests within the clusters leads to statistically insignificant correlations.

Overall Conclusion:
This leads to the conclusion that datasets with higher mean adversarial accuracy have lower mean transfer success rates, and vice versa. But within a single dataset, there is no statistically significant correlation between TSR and adversarial accuracy. Thus, these need to be studied separately.

---

4. Convergence with white-box attacks is inversely correlated with TSR

Convergence shall be measured as the inverse of accuracy drop i.e. $\frac{1}{accdrop}$

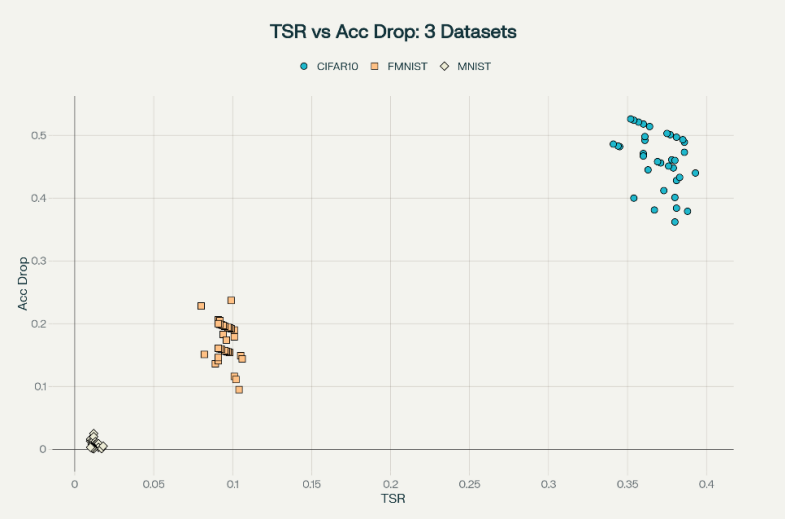
accuracy drop = accuracy Against transferred attack – accuracy against self (white-box) attack

Test Results:

- **Spearman's correlation coefficient (ρ): 0.866**
- **P-value: 1.12×10^{-33}**
- **Sample size: 108 pairs**

Interpretation:

- The correlation coefficient of **0. 866** indicates a **strong positive monotonic relationship** between acc_drop and TSR.
- The extremely low p-value confirms that this correlation is **statistically significant**, meaning the observed relationship is highly unlikely to be due to random chance.



Within each dataset, the correlations are either negative of insignificant.

Dataset	Spearman Correlation	p-value	Interpretation
cifar10	-0.417	0.011	Moderate negative correlation, statistically significant
fmnist	-0.238	0.163	Weak negative correlation, not statistically significant
mnist	-0.024	0.887	No meaningful correlation

This like (3) is another example of **Simpson's Paradox** (or the Yule-Simpson effect). It occurs when a trend appears in several different groups of data but reverses or disappears when these groups are combined. This implies the direct correlation between TSR and acc drop exits because of differences in datasets and within each dataset these trends are either negative of insignificant.

Overall Conclusion:

Within each dataset, convergence between transfer and self-attacks is directly related to TSR, since it is inversely related to accuracy drop. This correlation is only significant in CIFAR-10.

The strong correlation in the combined data implies datasets that have high transfer success rates (higher complexity) also have higher accuracy drops, that is white-box attacks dominate transfer attacks as dataset complexity increases, widening gap (reducing convergence).

---

5. Effect of entanglement strategy on Incoming and Outgoing TSR

Consider the data of classical -> Quantum transfer attacks (incoming).

- For CIFAR 10:

Lowest Mean TSR for CIFAR-10**: Full CZ – mean TSR (0.357)**

Statistical Significance:

A statistical test (independent t-test) comparing the TSR values for this pair against all other entg_strat, gate pairs yields:

- **t-statistic:** -5.30
- **p-value:** 0.017

Since the p-value is less than 0.05, the difference is **statistically significant**. This indicates that the full entg_strat with cz gate combination achieves a significantly lower mean TSR compared to other configurations for CIFAR-10.

- For FMNIST:

Lowest Mean TSR for FMNIST: **Linear CZ – mean TSR (0.081)**

Statistical Significance

A statistical test (independent t-test) comparing the TSR values for this pair against all other entg_strat, gate pairs yields:

- **t-statistic:** -8.83
- **p-value:** 0.000031

Since the p-value is much less than 0.05, the difference is **statistically significant**. This indicates that the linear entg_strat with cz gate combination achieves a significantly lower mean TSR compared to other configurations for FMNIST.

- For MNIST:

Lowest Mean TSR for MNIST: **Circular CNOT – mean TSR (0.010)**

Statistical Significance

A statistical test (independent t-test) comparing the TSR values for this pair against all other entg_strat, gate pairs yields:

- **t-statistic:** -6.44
- **p-value:** 0.000011

Since the p-value is much less than 0.05, the difference is **statistically significant**. This indicates that the circular entg_strat with cnot gate combination achieves a significantly lower mean TSR compared to other configurations for MNIST

➢ Aggregately:

Lowest Mean TSR Across All Datasets and Attacks: **Staggered CNOT – mean TSR (0.155)**

Statistical Significance

A statistical test (independent t-test) was conducted to compare the TSR values for this pair against all other entg_strat, gate pairs:

- **t-statistic:** -0.09
- **p-value:** 0.93

Since the p-value is much greater than 0.05, the difference is **not statistically significant**

Now consider the data of quantum -> classical TSR (outgoing)

➢ For CIFAR 10:

Highest Mean TSR for CIFAR-10**: Staggered CZ – mean TSR (0.3805)**
Statistical Significance

An independent t-test was performed to compare the TSR values for staggered, cz against all other entg_strat, gate pairs (excluding staggered, cz) for CIFAR-10:

- **t-statistic: 1.72**
- **p-value: 0.147**

Since the p-value is much greater than 0.05, the difference is **not statistically significant**

➢ For FMNIST:

Highest Mean TSR for FMNIST: **Circular CNOT – mean TSR (0.0995)**
Statistical Significance

A statistical test (independent t-test) was performed comparing the TSR values for this pair against all other entg_strat, gate pairs:

- **t-statistic:** 2.98
- **p-value:** 0.152

Since the p-value is greater than 0.05, the difference is **not statistically significant**.

- For MNIST:

Highest Mean TSR for MNIST: **Circular CZ – mean TSR (0.012)**
Statistical Significance

A statistical test (independent t-test) comparing the TSR values for this pair against all other entg_strat, gate pairs yields:

- **t-statistic:** 4.57
- **p-value:** 0.00037

Since the p-value is much less than 0.05, the difference is **statistically significant**. This indicates that the circular entg_strat with cz gate combination achieves a significantly higher mean TSR compared to other configurations for MNIST.

- For Aggregate:

Highest Mean TSR Across All Datasets and Attacks: **Full CNOT – mean TSR (0.151)**
Statistical Significance

A statistical test (independent t-test) was conducted to compare the TSR values for this pair against all other entg_strat, gate pairs:

- **t-statistic:** -0.10
- **p-value:** 0.92

Since the p-value is much greater than 0.05, the difference is **not statistically significant**

Results Matrix:

Dataset	Classical → Quantum	Quantum → Classical
CIFAR-10	Full CZ (significant)	Staggered CZ (insignificant)
FMNIST	Linear CZ (significant)	Circular CNOT (insignificant)
MNIST	Circular CNOT (significant)	Circular CZ (significant)
Aggregate	Staggered CNOT (insignificant)	Full CNOT (insignificant)

Overall Conclusion:

The results highlight the complex interplay between entanglement strategy, gate choice, dataset characteristics, and the direction of adversarial transferability. While certain entanglement strategies and gates can significantly impact TSR for specific attack scenarios and datasets, a universally optimal "entanglement strategy" that either consistently minimizes incoming TSR or maximizes outgoing TSR across all datasets and attack types is not supported by this data. This suggests that quantum-classical adversarial robustness and transferability are highly context-dependent, necessitating tailored approaches based on the specific dataset and the direction of the adversarial attack. Further research could explore the underlying reasons for these dataset-specific differences and investigate more sophisticated entanglement or gate strategies.