Robustness and Interpretability of Hybrid Quantum NLP Models

Anonymous ACL submission

Abstract

Hybrid quantum classical models offer theoretical advantages in expressivity and robustness, yet their practical utility in natural language processing (NLP) is still not well studied. This paper examines how variational quantum circuits behave when they applied to highly compressed text representations. A hybrid model is proposed where a frozen DistilBERT encoder converts each sentence into a fixed eight-dimensional representation. This compact representation is then passed to either a classical multilayer perceptron or a variational quantum head, with both options having a similar number of trainable parameters. To interpret these models, the paper defines *Interface Grad-CAM*, a mechanism that attributes importance at the shared interface and maps saliency back to tokens. On SST-2, AG News and Yelp Polarity, the quantum head consistently matches or slightly outperforms the classical head under the same eight dimensional bottleneck. More importantly, a *Quantum Shield* effect is observed: on SST-2, the synonym based attack success rate drops from about 47% for the classical head to about 17% for the quantum head, with a concurrent reduction on the other datasets. Gradient norm diagnostics at the interface indicate that this robustness does not arise from gradient masking. An entanglement analysis further reveals a modest negative correlation between global quantum entanglement and the entropy of token level importance scores, providing preliminary evidence that more highly entangled states may be associated with sharper, more focused explanations in the compressed feature space.

1 Introduction

Quantum machine learning is becoming a practical approach in which quantum approaches are used as a subroutine of a learning pipeline, instead of trying to replace the entire machine learning model with specialized quantum computers (Biamonte et al., 2017; Schuld and Killoran, 2019). Hybrid quantum classical neural networks combine quantum circuits with classical neural networks so that they can be trained together as one system. This approach has been tested on image classification and small structured datasets, where quantum parts work as learnable modules that can help improve performance when computing resources are limited (Farhi and Neven, 2018; Cong et al., 2019; Schuld et al., 2020).

In natural language processing, several works have coupled classical text encoders with quantum circuits for downstream classification tasks (Tacchino et al., 2020; Ardeshir-Larijani and Fatmehsari, 2024). These studies often show similar or slightly better accuracy when only a small amount of data is available, suggesting that quantum components might be able to capture useful patterns from compact text representations. However, there are two major limitations. First, many NLP hybrid models compare strong classical encoders with small quantum parts with much weaker classical models, making it hard to tell whether any improvement truly comes from the quantum component. Second, methods for explaining quantum models are much less developed than those for classical deep learning, especially for text tasks where understanding the importance of individual words is important for trust and interpretability (Ribeiro et al., 2016; Sundararajan et al., 2017). Quantum Grad-CAM methods for vision (Zhang et al., 2024) provide one direction to interpret variational circuits, but analogous techniques for hybrid text models remain rare.

This paper addresses these gaps by proposing and evaluating a hybrid quantum classical architecture centered on a shared small language model encoder and a carefully controlled low dimensional interface representation. The design is based on three simple ideas. First, the quantum and classical parts use the same input vector and have nearly

the same number of parameters. This makes sure that any difference in performance comes from the quantum processing itself, not from having a larger model. Second, explainability is a main focus. The model includes an Interface Grad-CAM method that traces the prediction back to the shared interface vector and the input words, creating clear importance maps that can be fairly compared for the quantum and classical parts. Third, the model studies robustness and entanglement. It checks how explanations change when words are replaced with similar ones and examines how the entanglement of the quantum circuit relates to the explanation patterns. The experiments use the SST-2 sentiment dataset with a strict eight-dimensional interface limit and are also tested on AG News and Yelp Polarity to see how the method behaves on other text classification tasks.

This paper is organized into different sections. The related work is discussed in Section 2, whereas methodology is explained in Section 3. Section 4 mentioned details about the experiments. Analysis of classical and quantum NLP models is provided in Section 5 and Section 6 concludes the findings with explanation about the future work.

2 Related Work

Research on quantum machine learning has established several foundations for hybrid quantum–classical architectures. Early work on quantum neural networks and quantum-inspired classifiers explored how variational quantum circuits can approximate nonlinear decision boundaries and be trained with gradient-based methods (Biamonte et al., 2017; Schuld and Killoran, 2019; Schuld et al., 2020). In vision and small-scale tabular tasks, studies on quantum classifiers and quantum convolutional networks demonstrated that quantum components can be embedded into classical models to achieve competitive performance under limited resources (Farhi and Neven, 2018; Cong et al., 2019; Tacchino et al., 2020).

Hybrid approaches for text classification have been investigated more recently. Some systems used fixed classical encoders to map sentences into continuous feature spaces, followed by quantum circuits that act as trainable classifiers (Tacchino et al., 2020). More advanced work on hybrid transfer learning showed that quantum heads attached to pretrained language models can match or slightly outperform classical heads when training data are scarce (Ardeshir-Larijani and Fatmehsari, 2024). However, these studies often compared quantum circuits against relatively shallow or under-parameterized classical baselines, making it difficult to disentangle the benefits of quantum structure from simple differences in model capacity.

Explainability is another critical dimension for modern NLP systems. Methods such as local surrogate explanations and gradient-based attribution have provided tools to interpret classical deep models (Ribeiro et al., 2016; Sundararajan et al., 2017). Grad-CAM and its extensions use gradients with respect to intermediate activations to produce class-specific saliency maps, offering an intuitive view of where a model "looks" when making decisions. Recently, quantum Grad-CAM variants have been proposed for visual applications (Zhang et al., 2024), but explainability for quantum text models remains relatively unexplored. In parallel, adversarial attacks on text classifiers, including synonym-based methods such as TextFooler (Jin et al., 2020), have highlighted the fragility of classical NLP models and motivated the study of robustness-aware explanations.

The present work connects these lines of research by combining a frozen language model encoder with matched quantum and classical heads, introducing an interface-level Grad-CAM mechanism, and evaluating both robustness and entanglement-based properties of quantum explanations.

3 Methodology

The architecture is designed for sentence level classification tasks. The empirical analysis in this paper focuses on three benchmarks: SST-2, AG News and Yelp Polarity.

3.1 Datasets Description

The Stanford Sentiment Treebank (SST-2) is a binary sentiment classification dataset derived from the Stanford Sentiment Treebank (Socher et al., 2013). Each example consists of a single English sentence and a label $y \in \{0, 1\}$, where 0 denotes negative sentiment and 1 denotes positive sentiment. The training set contains approximately 67 000 examples, and the validation set contains 872 examples. The gold labels for the official test split are hidden, so this paper reports validation metrics and uses the test split primarily for explain-

ability and robustness analysis.

AG News is a dataset used for classifying news into four categories: World, Sports, Business, and Science/Technology, based on news titles and short descriptions (Zhang et al., 2015). Yelp Polarity is a large-scale binary sentiment dataset of user reviews in which labels indicate overall positive or negative sentiment (Zhang et al., 2015). For all datasets, a held out portion of the original training data is used for validation, and the official test sets are used for reporting aggregate performance and robustness metrics.

3.2 Preprocessing and Tokenization

Each input sentence is first processed by the DistilBERT tokenizer, which splits the text into subword tokens drawn from a fixed vocabulary. Sequences longer than a maximum length L are truncated, and shorter sequences are padded to length L. An associated binary attention mask marks real tokens and padding positions. The model receives token identifiers and the corresponding attention mask as input. The shared encoder is a frozen DistilBERT model that maps the token sequence and the attention mask to contextualized hidden states. These token representations are mean pooled using the attention mask to obtain a single sentence vector, which is then passed through a learned linear projection to form the interface vector $\mathbf{z} \in \mathbb{R}^D$. In the experiments, the dimension of the interface is set to $D = 8$. Figure 1 illustrates the data flow from raw text through tokenization, DistilBERT, mean pooling and projection into the shared interface.

3.3 Classical and Quantum Variational Head

The classical head is a lightweight MLP operating on the interface vector $\mathbf{z} \in \mathbb{R}^D$: LayerNorm→ReLU→FC($D\rightarrow$ 16) →ReLU→FC(16→ K).

The quantum head treats the interface vector $\mathbf{z} \in \mathbb{R}^D$ as a set of input angles to a variational quantum circuit with D qubits. The circuit uses Pennylane's `StronglyEntanglingLayers` template, which consists of repeated layers of parameterized single qubit rotations and entangling two qubit gates.

Let, $n = D$ denote the number of qubits and let L denote the number of entangling layers. The quantum circuit performs two main steps: data embedding, which applies rotations based on the components of $\mathbf{z}$, and variational evolution, which applies L layers of parameterized entangling gates with trainable angles $\boldsymbol{\theta}$. The data embedding unitary can be written abstractly as

$$U_{\text{embed}}(\mathbf{z}) = \prod_{i=1}^{n} R(\mathbf{z}_i, i), \tag{1}$$

where $R(\mathbf{z}_i, i)$ denotes a composition of single-qubit rotations on qubit i with angles derived from $\mathbf{z}_i$. The variational part applies L layers of parameterized gates:

$$U_{\text{var}}(\boldsymbol{\theta}) = \prod_{\ell=1}^{L} U_\ell(\boldsymbol{\theta}_\ell), \tag{2}$$

where each U_ℓ couples all qubits through a prescribed entangling topology and $\boldsymbol{\theta}_\ell$ collects the trainable rotation angles in layer ℓ. The full unitary is therefore

$$U(\mathbf{z}, \boldsymbol{\theta}) = U_{\text{var}}(\boldsymbol{\theta})\, U_{\text{embed}}(\mathbf{z}). \tag{3}$$

Starting from the all-zero state $|0\rangle^{\otimes n}$, the circuit prepares

$$|\psi(\mathbf{z}, \boldsymbol{\theta})\rangle = U(\mathbf{z}, \boldsymbol{\theta})|0\rangle^{\otimes n}. \tag{4}$$

The readout measures the expectation value of the Pauli Z operator on each qubit:

$$q_i(\mathbf{z}, \boldsymbol{\theta}) = \langle \psi(\mathbf{z}, \boldsymbol{\theta})| Z_i |\psi(\mathbf{z}, \boldsymbol{\theta})\rangle, \tag{5}$$

and assembles these into a feature vector

$$\mathbf{q}(\mathbf{z}, \boldsymbol{\theta}) = (q_1, \ldots, q_n) \in \mathbb{R}^n. \tag{6}$$

This vector is then passed through a small classical post-processing network, mirroring the structure of the classical head, to produce logits over the two sentiment labels. Figure 2 illustrates the data flow inside the quantum head.

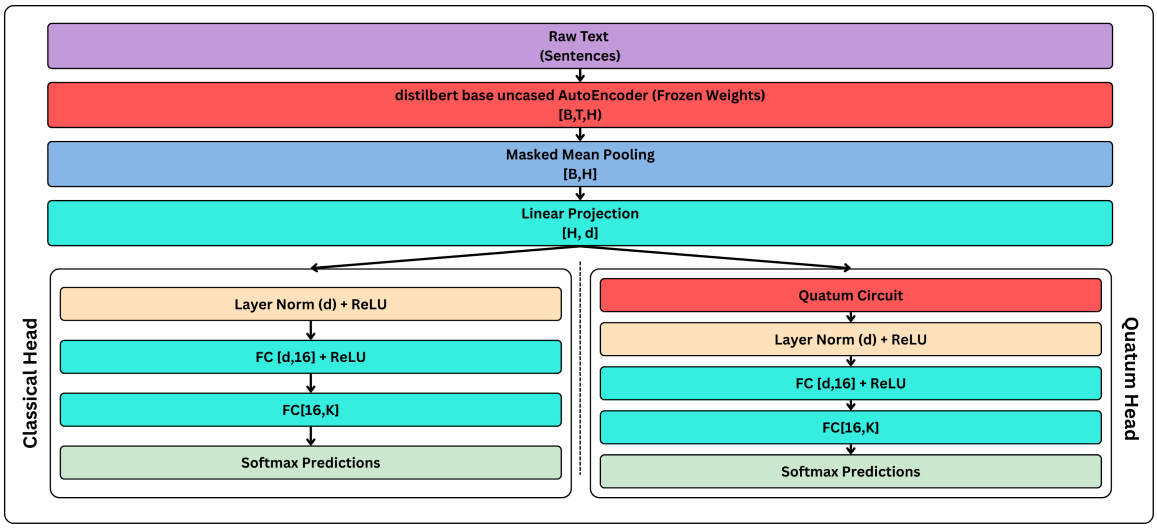

Figure 1: Hybrid quantum–classical architecture used in this work. Raw text is encoded by a frozen DistilBERT encoder, pooled, and linearly projected to a d-dimensional interface that feeds either a classical head or a quantum head to produce class logits and softmax predictions. Notation: B is batch size, T is sequence length, H is encoder hidden size, d is interface dimension (and number of qubits), K is number of classes and FC denotes a fully connected (linear) layer.

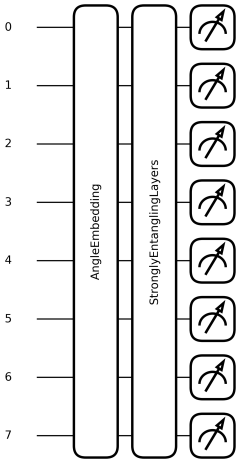

Figure 2: PennyLane circuit for the quantum head: `AngleEmbedding` encodes $\mathbf{z} \in \mathbb{R}^d$ on d qubits, followed by `StronglyEntanglingLayers` (L=4), with $\langle Z \rangle$ readout on each qubit producing $\mathbf{q} \in \mathbb{R}^d$.

3.4 Loss Function and Optimization

For a given example (x, y) with label $y \in \{0, 1\}$, each head (classical or quantum) produces logits that are converted to class probabilities with a softmax layer. Training minimizes the standard cross-entropy loss, which penalizes low probability assigned to the true label and encourages confident correct predictions. This choice is convenient for the quantum head because gradients can be propagated through the hybrid graph using parameter shift rules for the variational circuit (Schuld et al., 2019).

Training is performed using AdamW optimizer (Loshchilov and Hutter, 2019). A smaller learning rate is applied to the quantum circuit parameters than to the projection and MLP layers, reflecting empirical observations that conservative updates help avoid barren plateaus (McClean et al., 2018). Gradient clipping is used to prevent very large updates and keep training stable. Training stops early if the validation loss stops improving, and the learning rate is reduced automatically when progress slows down.

3.5 Interface Grad-CAM for Hybrid Models

To interpret the predictions, the architecture uses an interface Grad-CAM mechanism that adapts the original Grad-CAM (Selvaraju et al., 2017) approach to the shared interface representation. Given the interface vector $\mathbf{z}$ and logits $\boldsymbol{\ell}$, the explanation focuses on the logit associated with the predicted class c:

$$\ell_c = \boldsymbol{\ell}_c. \tag{7}$$

The gradient of ℓ_c with respect to $\mathbf{z}$ is computed:

$$\mathbf{g} = \nabla_{\mathbf{z}} \ell_c. \tag{8}$$

The interface saliency vector $\mathbf{s}^{\text{int}} \in \mathbb{R}^D$ is defined by elementwise gradient–activation products:

$$\mathbf{s}^{\text{int}} = \mathbf{z} \odot \mathbf{g}, \tag{9}$$

where $\odot$ denotes element wise multiplication. The saliency is then normalized to the interval $[0, 1]$:

$$\mathbf{s}_{\text{norm}}^{\text{int}} = \frac{\mathbf{s}^{\text{int}} - \min_j s_j^{\text{int}}}{\max_j s_j^{\text{int}} - \min_j s_j^{\text{int}} + 10^{-8}}. \quad (10)$$

To obtain token level scores, the method uses the encoder projection weights. Recall that

$$\mathbf{z} = \mathbf{W}_{\text{proj}}\mathbf{h}_{\text{pool}} + \mathbf{b}_{\text{proj}}. \quad (11)$$

The interface weights can be used to define a vector in the encoder hidden space:

$$\mathbf{v} = \mathbf{s}_{\text{norm}}^{\text{int}\,\top}\mathbf{W}_{\text{proj}}, \quad (12)$$

where $\mathbf{v} \in \mathbb{R}^H$ represents a global importance direction. For each token position i, a raw token importance score is computed as:

$$r_i = \mathbf{h}_i^\top \mathbf{v}. \quad (13)$$

The scores are masked and normalized over non-padding positions to produce a probability distribution over tokens:

$$p_i^{\text{tok}} = \frac{\max(r_i, 0) \cdot m_i}{\sum_{j=1}^{L} \max(r_j, 0) \cdot m_j + 10^{-12}}. \quad (14)$$

These token level probabilities form the basis for the entropy, stability and faithfulness metrics described in the next section.

4 Experiments

4.1 Experimental Setup

All experiments use SST-2 with a training subset of 20 000 sentences randomly sampled from the official training set. The full development set of 872 examples is used for validation. DistilBERT is frozen and used as a feature extractor with maximum sequence length $L = 128$. The interface dimension is $D = 8$ for both heads. The optimizer is AdamW with base learning rate 10^{-3}, and the quantum circuit parameters use a scaled learning rate of 3×10^{-4}. Batch size is 32. The main baselines consist of a classical head attached to the shared interface and quantum head as described in Section 3.3.

4.2 Training Dynamics

On SST-2, the classical head converges quickly in about four epochs, reaching approximately 0.79 accuracy with validation loss dropping from about 0.65 to 0.46 before plateauing. The quantum head starts slower but continues improving, ultimately achieving around 0.83 accuracy with validation loss decreasing from about 0.70 to 0.38 before early stopping.

Model	Val Acc	Val Loss
Classical Head	0.79	0.46
Quantum Head	**0.83**	**0.38**

Table 1: Validation accuracy and loss on SST-2. The hybrid heads share the interface vector and have closely matched parameter counts.

Dataset	Model	Val Acc	Val Loss
SST-2	Classical	0.79	0.46
	Quantum	**0.83**	**0.38**
AG News	Classical	0.88	0.33
	Quantum	**0.89**	0.33
Yelp Polarity	Classical	0.86	0.33
	Quantum	**0.88**	**0.29**

Table 2: Validation accuracy and loss across datasets under a shared eight dimensional representation.

Gradient clipping and differential learning rates are essential for stable training of the quantum head. Without clipping, large gradient spikes occasionally occur and can destabilize training. Constraining the gradient norm yields smoother updates, and using a smaller learning rate for quantum parameters improves convergence.

Table 1 summarizes the main validation results on SST-2, and Table 2 reports validation accuracy and loss across datasets under the same eight-dimensional interface. The quantum head improves validation accuracy and loss on SST-2 while using the same frozen encoder and interface representation. This suggests that the quantum circuit can exploit the compact interface representation more effectively than a shallow MLP when appropriately tuned. The total number of trainable parameters differs by fewer than 100 parameters between the classical and quantum variants, and both occupy approximately 253 MB on disk. On a CPU-only setup, a full training run on SST-2 takes about 14 minutes for the classical head and about 19 minutes for the quantum head, reflecting the additional cost of quantum circuit simulation.

4.3 Additional Datasets

To assess whether the observations on SST-2 generalize to other text classification tasks, the same frozen DistilBERT encoder was evaluated on two additional datasets: AG News (four-way topic clas-

sification) and Yelp Polarity (binary review sentiment). For each dataset, the encoder is frozen, the interface dimension is fixed at $D = 8$ for both heads and the models are trained under the same optimization protocol with $10\,000$ training examples, $10\,000$ validation examples and a $4\,000$-example test subset. On AG News, the classical head achieves validation and test accuracies of approximately 0.88 and 0.89 respectively, while the quantum head reaches very similar values (validation accuracy 0.89, test accuracy 0.89) with a lower synonym-based attack success rate (0.09 vs. 0.16). On Yelp Polarity, the classical head achieves validation accuracy around 0.86 (test accuracy 0.87), and the quantum head performs slightly better, with validation accuracy about 0.88 and test accuracy about 0.88 and a lower adversarial success rate (0.03 vs. 0.09). Across all three datasets (SST-2, AG News and Yelp Polarity), the quantum head therefore matches or slightly surpasses the classical head in accuracy under the same eight-dimensional interface constraint and consistently exhibits lower synonym-based attack success rates, although the absolute margins remain modest and dataset dependent.

4.4 Ablation Studies

Several ablation studies quantify the contributions of design choices.

Interface dimension: Reducing D from 8 to 4 degrades the quantum head by roughly two points, while the classical head is less affected. This suggests that the quantum circuit benefits from a slightly richer interface, likely because fewer qubits limit the entanglement structure.

Quantum depth and readout: Using only $L = 2$ entangling layers and measuring a single qubit yields lower accuracy and weaker robustness. In contrast, the multi-qubit readout with $L = 4$ layers provides better performance, confirming that deeper entanglement and richer readouts are beneficial in the bottleneck setting.

Learning-rate schedule: Using a smaller learning rate for quantum parameters improves training stability and final accuracy. Training the quantum head with the same learning rate as the classical head produces noisier validation curves and slightly lower final accuracy. The smaller learning rate for quantum parameters combined with ReduceLROnPlateau yields smoother convergence and higher accuracy.

Gradient clipping: Clipping gradients (max norm 1.0) prevents occasional large updates and stabilizes training. Disabling clipping can lead to validation degradation and occasional divergence. Clipping with a norm of 1.0 significantly stabilizes training.

5 Analysis

The classification reports generated on the validation set provide a detailed view of per class performance. The classical head exhibits strong precision on negative examples but lower recall, indicating a tendency to favor positive predictions. The quantum head is more balanced across classes, with f1-scores above 0.82, reducing both error types. The confusion matrix statistics confirm that the quantum model reduces both types of errors.

5.1 Explainability Metrics

Using the token-level saliency distributions p_i^{tok}, the following XAI metrics are computed.

Entropy: The normalized entropy of the saliency distribution is

$$H = -\frac{1}{\log N} \sum_{i=1}^{N} p_i^{\text{tok}} \log \left(p_i^{\text{tok}} + 10^{-12} \right), \quad (15)$$

where N is the number of valid tokens. Lower H indicates more concentrated explanations. Both heads produce reasonably focused explanations, with small dataset-dependent differences (Table 3). For example, on SST-2 the classical head achieves an entropy of 0.940 while the quantum head achieves 0.937.

Stability: To measure stability under perturbations, random subsets of tokens are masked in the input, and new saliency distributions are computed. Stability is defined as the cosine similarity between the original and perturbed saliency vectors:

$$S = \frac{\sum_i p_i^{\text{tok}} q_i^{\text{tok}}}{\sqrt{\sum_i (p_i^{\text{tok}})^2} \sqrt{\sum_i (q_i^{\text{tok}})^2}}, \quad (16)$$

where $\mathbf{q}^{\text{tok}}$ is the perturbed distribution. The quantum head achieves slightly higher mean stability than the classical head.

Faithfulness: To evaluate faithfulness, the top-k salient tokens are masked, and the drop in predicted probability for the originally predicted class is measured:

$$\Delta = p_c^{\text{orig}} - p_c^{\text{masked}}, \quad (17)$$

where p_c^{orig} is the original predicted probability and p_c^{masked} is the probability after masking. Larger Δ

Dataset	Model	Entropy ↓	Stability ↑	Faith. drop ↑
SST-2	Classical	0.940	0.922	0.478
	Quantum	**0.937**	**0.932**	**0.304**
AG News	Classical	0.973	0.954	0.369
	Quantum	**0.964**	**0.946**	**0.196**
Yelp Polarity	Classical	0.977	**0.958**	**0.425**
	Quantum	0.977	0.939	0.164

Table 3: Token level explainability metrics at the interface for each dataset. Entropy measures concentration of saliency (lower is more focused), stability measures robustness of token scores under perturbations (higher is better) and the probability drop after masking salient tokens measures faithfulness (higher indicates that salient tokens are influential).

Dataset	Model	Attack succ. ↓	Saliency sim. ↑
SST-2	Classical	0.469	0.693
	Quantum	**0.172**	**0.794**
AG News	Classical	0.156	0.812
	Quantum	**0.094**	**0.915**
Yelp Polarity	Classical	0.094	**0.861**
	Quantum	**0.031**	0.856

Table 4: Adversarial robustness metrics across datasets. Attack success is the fraction of synonym-based adversarial examples that flip the prediction (lower is better), and saliency similarity is the cosine similarity between Grad-CAM token scores for clean and adversarial inputs (higher indicates more stable explanations).

indicates that the explanation identifies truly influential tokens. The classical head shows a larger drop on average, whereas the quantum head has a somewhat smaller but still substantial drop, suggesting a more distributed yet meaningful attribution.

5.2 Qualitative Grad-CAM Visualizations

In addition to aggregate metrics, the Grad-CAM mechanism produces token-level heatmaps for individual sentences. For representative examples where both heads predict the same label, the classical head often concentrates on a small set of strongly polarized tokens, whereas the quantum head tends to distribute saliency across sentiment-bearing phrases and contextual modifiers. This qualitative difference is consistent with the slightly higher stability of the quantum head's saliency distribution: the quantum heatmap changes less under small perturbations.

5.3 Quantum Shield

The Quantum Shield analysis uses a synonym-based adversarial attack inspired by TextFooler (Jin et al., 2020). For each input sentence, salient tokens are identified via interface Grad-CAM, and candidate replacements are proposed using a masked language model. The attack attempts to alter the model's prediction while preserving approximate semantics. Across datasets, the quantum head consistently exhibits lower or comparable attack success rates and equal or higher saliency similarity between clean and adversarial examples than the classical head. On SST-2, for example, the attack success rate drops from about 47% for the classical head to about 17% for the quantum head. Moreover, the quantum head typically retains higher cosine similarity between Grad-CAM token scores for clean and adversarial inputs, indicating more stable explanations under synonym perturbations.

Table 4 summarizes adversarial robustness metrics across datasets.

Figure 4 visualizes the Quantum Shield results across datasets.

5.4 Gradient Norms at the Interface

To verify that the observed robustness is not an artifact of gradient masking, we measure the L2 norm of $\nabla_{\mathbf{z}} \log p(\hat{y} \mid x)$: it is about 0.43 (classical) versus 0.63 (quantum), indicating comparable interface sensitivity. This suggests that the lower attack success rate of the quantum head is not due to vanishing gradients but reflects genuinely different behavior under synonym-based perturbations.

5.5 Entanglement as Attention

To investigate the relationship between entanglement and explanations, the Meyer Wallach global entanglement measure is computed for the quantum state $|\psi(\mathbf{z}, \boldsymbol{\theta})\rangle$. For each qubit i, the reduced density matrix ρ_i is obtained by tracing out all other qubits. The measure is

$$Q(|\psi\rangle) = \frac{2}{n} \sum_{i=1}^{n} \left(1 - \mathrm{Tr}\left(\rho_i^2\right)\right). \qquad (18)$$

Across validation samples, the quantum head produces highly entangled states (mean $Q \approx 0.99$), and Q correlates negatively with saliency entropy (≈ -0.27), suggesting that higher entanglement is associated with sharper token-level explanations.

6 Conclusion

This paper presents a hybrid language model that combines a frozen DistilBERT encoder with either a classical or a quantum classifier, both using the

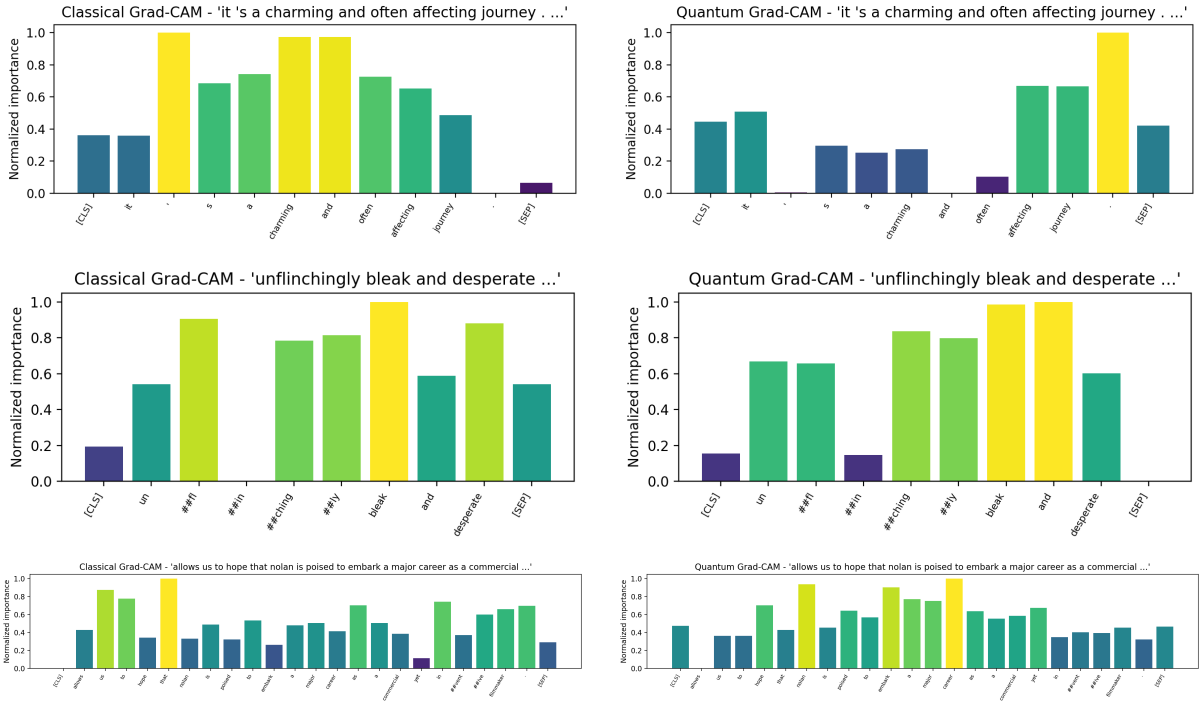

Figure 3: Representation of Grad-CAM token saliency visualizations for the classical and quantum heads.

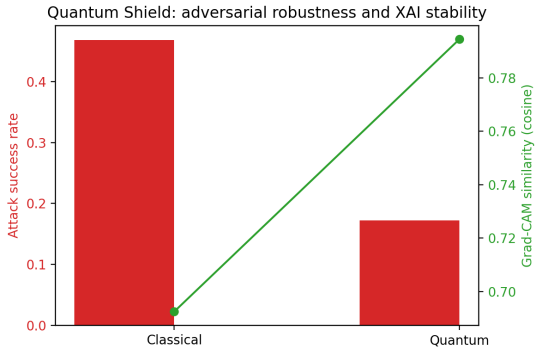

Figure 4: Quantum Shield Results

same shared representation. The model is designed so that the classical and quantum parts have a similar number of parameters, allowing a fair comparison. It also includes an Interface Grad-CAM method to explain predictions at the word level and studies model robustness and quantum entanglement. On the SST-2 dataset, the quantum classifier performs better than the classical one, achieving higher validation accuracy and lower loss. It is also much more robust to attacks where words are replaced with similar ones. The entanglement analysis shows that highly entangled quantum states are linked to clearer and more focused explanations, suggesting that entanglement plays a useful role in shaping how the model explains its decisions.

The overall findings suggest that hybrid quantum–classical architectures can serve not only as accuracy-competitive models but also as testbeds for studying the interplay between quantum properties such as entanglement and classical notions of explanation and robustness. Future work can extend the experiments to additional datasets, hardware implementations and tasks such as natural language inference or question answering, as well as explore circuits explicitly designed to optimize both performance and explanation quality.

Limitations

The present study is limited in several ways. First, the full explanation analyzes (interface Grad-CAM, robustness and entanglement) are conducted on SST-2; the AG News and Yelp Polarity experiments provide additional evidence under the same shared-interface design but are restricted to three medium-scale classification tasks and do not cover the wider variety of NLP problems such as natural language inference or question answering. Second, the quantum circuits are simulated on classical hardware; this leads to substantially higher training and inference times than shallow classical heads, and results on actual quantum devices may differ due to noise and hardware-specific limitations. Third, the comparison between quantum and

classical heads focuses on matched parameter budgets and a single encoder; alternative architectures or larger encoders may yield different relative performance. Fourth, the entanglement-as-attention correlation is demonstrated empirically but does not constitute a causal proof; further theoretical analysis is required.

References

Farzaneh Ardeshir-Larijani and Mehran Fatmehsari. 2024. Hybrid classical–quantum transfer learning for text classification. *Quantum Machine Intelligence*.

Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. 2017. Quantum machine learning. *Nature*, 549(7671):195–202.

Iris Cong, Soonwon Choi, and Mikhail D Lukin. 2019. Quantum convolutional neural networks. In *Nature Physics*, volume 15, pages 1273–1278.

Edward Farhi and Hartmut Neven. 2018. Classification with quantum neural networks on near term processors. *arXiv preprint arXiv:1802.06002*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8018–8025.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations*.

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

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. 2019. Evaluating analytic gradients on quantum hardware. *Physical Review A*, 99(3):032331.

Maria Schuld, Alex Bocharov, Krysta M Svore, and Nathan Wiebe. 2020. Circuit-centric quantum classifiers. *Physical Review A*, 101(3):032308.

Maria Schuld and Nathan Killoran. 2019. Quantum machine learning in feature hilbert spaces. *Physical Review Letters*, 122(4):040504.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328.

Francesco Tacchino, Chiara Macchiavello, Dario Gerace, and Daniele Bajoni. 2020. Quantum implementation of an artificial feed-forward neural network. *Quantum Science and Technology*, 5(4):044010.

Wei Zhang and 1 others. 2024. Quantum gradient-based class activation mapping for model interpretability. *arXiv preprint arXiv:2408.05899*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 28.