nature communications

Article

https://doi.org/10.1038/s41467-026-70420-4

Superior resilience to poisoning and amenability to unlearning in quantum machine learning

Received: 16 September 2025

Accepted: 26 February 2026

Published online: 09 March 2026

Check for updates

Yu-Qin Chen [1] & Shi-Xin Zhang [2]

The reliability of artificial intelligence hinges on the integrity of its training data, a foundation often compromised by noise and corruption. Here, through a comparative study of classical and quantum neural networks on both classical and quantum data, we reveal a fundamental difference in their response to data corruption. We find that classical models exhibit brittle memorization, leading to a failure in generalization. In contrast, quantum models demonstrate superior resilience, underscored by a phase transition-like response to increasing noise, revealing a critical point beyond which the model's performance changes qualitatively. We further introduce a framework of quantum machine unlearning, the process of efficiently forcing a trained model to forget bad influences. We show that classical models form rigid, stubborn memories of erroneous data, while the quantum model is significantly more amenable to efficient forgetting with approximate unlearning methods. Our findings establish that quantum machine learning possesses the dual advantage of intrinsic resilience and efficient adaptability, providing a promising paradigm for the trustworthy and reliable artificial intelligence of the future.

The remarkable success of artificial intelligence (AI) has placed it as a transformative force in science and society[1–3], yet the reliability of these powerful models is fundamentally contingent on the integrity of their training data. In real-world applications, datasets are often compromised by corruption, such as mislabeled examples or malicious poisoning attacks, which can severely impair a model's generalization, introduce dangerous biases, and create significant security vulnerabilities[4–8]. A closely related and increasingly urgent challenge is that of machine unlearning: the need to efficiently remove the influence of specific data from a trained model to comply with privacy regulations or to correct for erroneous information[9–15]. While retraining a model from scratch on a sanitized dataset offers a definitive solution, its prohibitive computational cost for large-scale systems makes it impractical, driving the search for more efficient alternatives.

As the field of quantum computing evolves rapidly[16–18], quantum machine learning (QML) has emerged as a promising paradigm with the potential to solve problems intractable for classical computers[19–30]. These models, often implemented as quantum neural networks (QNNs), leverage principles like superposition and entanglement to navigate vast computational spaces[31–48]. However, despite rapid theoretical and experimental progress, the behavior of quantum models in the face of real-world data imperfections remains a critical and largely unexplored frontier. While some studies have investigated the vulnerability of QNNs to inference-time adversarial attacks[49–62], their response to training-time data corruption is poorly understood. This distinction is critical: adversarial examples challenge a model's perception at inference, whereas data poisoning corrupts the model's fundamental knowledge acquisition. Furthermore, the subsequent problem of how to correct a poisoned model, a field we establish here as quantum machine unlearning, remains an entirely unexplored area. This gap presents a formidable barrier to developing quantum AI that is not only powerful but also trustworthy and secure.

[1]Graduate School of China Academy of Engineering Physics, Beijing, China. [2]Institute of Physics, Chinese Academy of Sciences, Beijing, China.
e-mail: yqchen@gscaep.ac.cn; shixinzhang@iphy.ac.cn

In this work, we bridge these critical gaps through a systematic, comparative study of a classical machine learning model of multi-layer perceptron (MLP) and a QNN, evaluating their response to different data corruption and different unlearning approaches across both classical and quantum datasets (Fig. 1). We reveal a fundamental difference in their learning mechanisms. The classical model exhibits brittle memorization, attempting to fit all data points, which leads to a catastrophic collapse in generalization when faced with contradictory information. In stark contrast, the QNN demonstrates a intrinsic resilience, prioritizing the general data structure over statistical outliers. This robust behavior is underscored by a distinct phase transition-like response to increasing label noise. This suggests that the QNN's learning dynamics is not a simple fitting exercise but a complex system exhibiting critical phenomena, a behavior far more structured than the continuous performance degradation of the MLP. Building on this discovery, we investigate the field of quantum machine unlearning. We find that the brittle nature of the classical model forms deep, stubborn memories that make efficient unlearning profoundly challenging. Conversely, the quantum model displays plasticity, proving highly amenable to efficient forgetting. Approximate unlearning methods are not only more stable but can achieve better performance than retraining from scratch within the same training time window. This work systematically demonstrates the dual advantage for QML by providing both a foundational analysis of controlled levels of data corruption in QML and the quantum machine unlearning framework addressing a previously unexplored aspect of quantum model forgetting, thereby paving the way for the development of secure and adaptable quantum technologies.

Results

To systematically investigate the behavior of machine learning models in corrupted environments, we conducted a comparative study centered on a classical MLP and a QNN built on top of parameterized quantum circuits (PQC). We focused on data poisoning in supervised learning paradigm and hence these models were evaluated on two distinct binary classification tasks (Fig. 2b): a classical task using MNIST handwritten digits ('1' vs '9') and a quantum task classifying the ground state phases of the one-dimensional XXZ spin Hamiltonian. We first probe the intrinsic resilience of each model by corrupting a controlled fraction α of the training data using two distinct protocols: Label Flipping and Feature Randomization (Fig. 1b). The former introduces direct contradictions by inverting a sample's class label, while the latter destroys the data's inherent structure by replacing the entire feature vector with random noise. Subsequently, we assess each model's capacity for repair by applying four different machine unlearning algorithms to remove the influence of the corrupted data. The following results reveal a consistent dual advantage for the quantum model, demonstrating superior resilience to training data corruption and greater amenability to efficient forgetting.

Superior resilience of quantum models to data corruption

In the Label Flipping scenario, where a fraction α of training labels are inverted, the MLP and QNN exhibit starkly different resilience behaviors (Fig. 2c, d), revealing fundamental differences in their ability to distinguish signal from noise.

The classical MLP shows a high sensitivity to label noise, resulting in a continuous compromise of its generalization ability. As shown in Fig. 2c, d, its validation accuracy begins to degrade almost immediately and steadily with any increase in the noise ratio α. This behavior stems from its tendency towards brittle memorization as illustrated in Fig. 2a. MLP attempts to accommodate every contradictory data point, causing a steady erosion of the global decision boundary's integrity. While its training accuracy remains high, this comes at the direct cost of its performance on unseen data, indicating an inability to effectively disregard even low levels of noise.

In stark contrast, the QNN displays a robust strategy characterized by a resilient performance plateau followed by a sharp, critical transition at $\alpha = 0.5$. As shown in Fig. 2c, d, the system exhibits two distinct phases: a signal-dominated phase for $\alpha < 0.5$, where validation accuracy remains high, and a noise-dominated phase for $\alpha > 0.5$, where performance collapses. Within the signal-dominated phase, the QNN effectively ignores the noisy outliers–a behavior confirmed by its decreasing training accuracy–thereby preserving the integrity of its learned representation. The sharp drop at the critical point $\alpha_c \approx 0.5$ is characteristic of a phase transition. This behavior is the hallmark of a robust system that maintains its ordered state (correct classification)

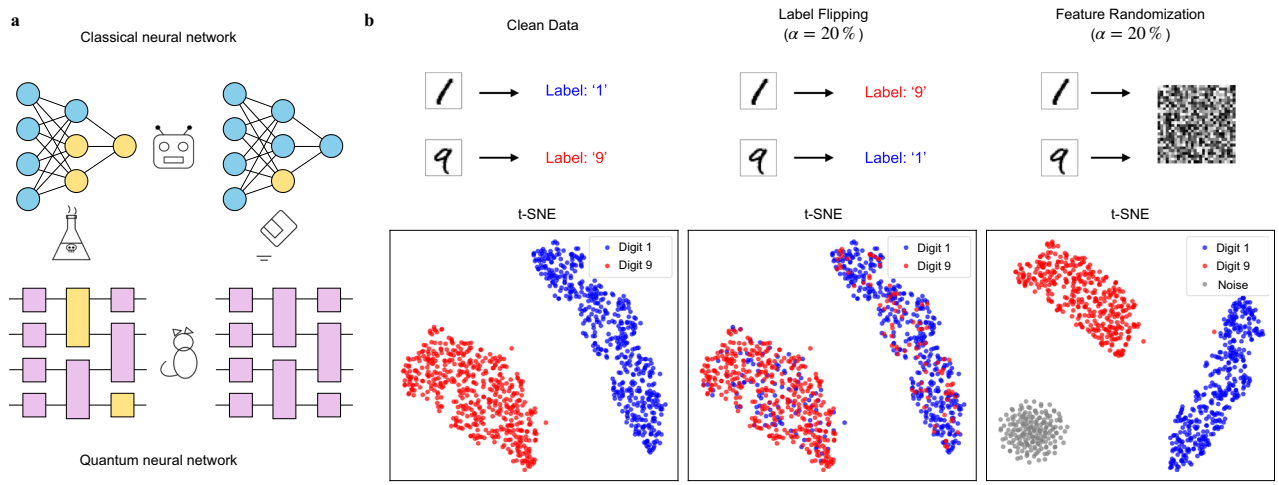

Fig. 1 | Conceptual framework and data corruption protocols. a Conceptual illustration of the learning and unlearning lifecycle investigated in this work. Both the classical neural network and the quantum neural network are subjected to training-time data corruption (poisoning) and subsequent correction (machine unlearning). The distinct icons for each model represent their different computational paradigms. QNN shows superior performance for learning under data corruption and unlearning. **b** Visualization of the data corruption protocols on the MNIST dataset, illustrated with representative examples and t-SNE low-dimension projections of the feature space. The leftmost column shows the clean data, where the t-SNE plot reveals two well-separated clusters for digits `1' (blue) and `9' (red). The middle column illustrates Label Flipping with a noise ratio of $\alpha = 0.2$, creating in-cluster anomalies as points retain their geometric position but adopt the opposing class's label. The rightmost column shows Feature Randomization ($\alpha = 0.2$), which replaces input images with random vectors, introducing a new, structurally distinct cluster of noise points (gray).

https://doi.org/10.1038/s41467-026-70420-4

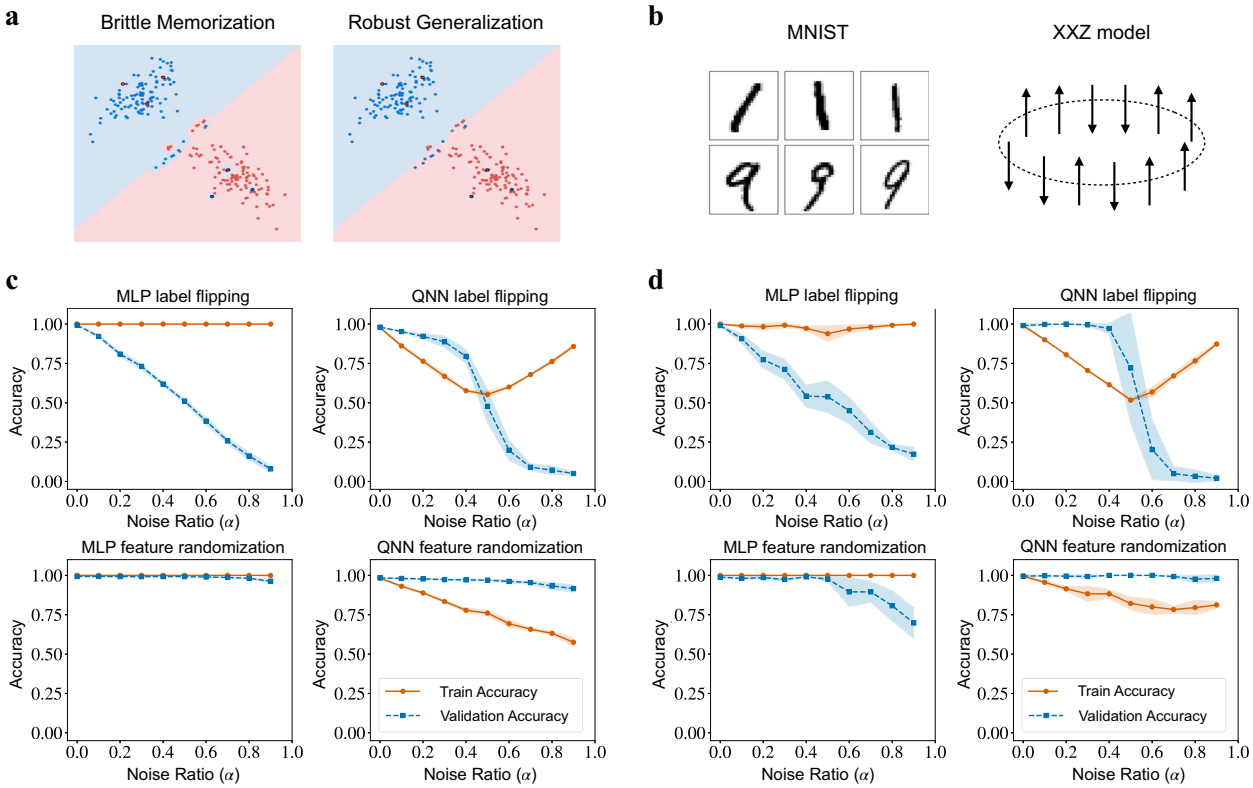

Fig. 2 | Superior resilience of quantum models to data corruption. a Conceptual illustration of the two learning paradigms. MLP multi-layer perceptron, QNN quantum neural network. The classical MLP engages in brittle memorization, deforming its decision boundary to fit mislabeled outliers (e.g., blue dots in red region). The QNN exhibits robust generalization, maintaining a simple boundary and correctly classifying the majority of points. Note that the decision boundary is only for illustration purpose and is not strictly from real experiments. **b** Representative data from the two classification tasks. Distinguishing MNIST digits `1' and `9' and distinguishing ground state phases of the 1D XXZ spin model. **c** Empirical results for the classical MNIST dataset. Plots show training (orange solid) and validation (blue dashed) accuracy versus the data noise ratio α. Under label flipping, the MLP's validation accuracy continuously degrades, while the QNN maintains a robust performance plateau before a sharp critical transition. Both models are robust to feature randomization with moderate α. Shaded regions represent the standard deviation over 5 optimization runs. **d** Empirical results for the quantum XXZ dataset. The trends are consistent with the quantum data, showing the MLP's continuous compromise versus the QNN's robust plateau under label flipping. The QNN's resilience is thus a general feature, independent of the nature of the data.

against disorder (label noise) up to a definitive transition point. In other words, the label noise is a relevant perturbation for classical MLP but an irrelevant perturbation for QNN. This advantage is not merely about model size, as it persists even when comparing QNNs of varying depth against MLPs of varying capacity (see Supplementary Information Section III). The ability of QNN to maintain a predictable window of high performance under a significant degree of contamination makes it an inherently more reliable model for real-world applications where training data is inevitably imperfect.

To ensure that the observed brittle memorization is universal to classical learning paradigms rather than a result of specific model architecture or capacity, we extended our evaluation to include more models ranging from linear classifiers such as support vector machine and logistic regression to modern architectures such as convolutional neural networks, ResNet, and random Fourier feature networks (see Supplementary Information Section II and IV). On the low-capacity end, simple linear classifiers exhibit immediate linear degradation, indicating that standard convex optimization forces the decision boundary to shift to accommodate noise even without over-parameterization. On the high-capacity end, modern deep architectures with increased depth, width, and features like residual connections, batch normalization, and low frequency filters, all consistently exhibit the continuous compromise behavior under label noise. These results confirm that the phase-transition-like resilience is a unique advantage stemming from the unitary structure of the quantum model.

In the Feature Randomization scenario, where input feature vectors are replaced with random noise, both models exhibit considerable robustness, as the out-of-distribution nature of the corrupted data is generally easy to distinguish[63]. However, a closer examination, particularly in data-limited regimes, reveals a more nuanced picture and a distinct advantage for the QNN.

On the smaller XXZ dataset, the MLP's performance degrades notably at high corruption ratios ($\alpha > 0.8$), where the number of remaining clean samples becomes low (Fig. 2d). The model struggles to form a stable decision boundary amidst a sea of random noise. In stark contrast, the QNN's validation accuracy remains almost entirely unaffected across all values of α (Fig. 2d). This indicates a better generalization ability for QML models[64] to identify and completely disregard out-of-distribution noise, even when clean training examples are very scarce.

This performance difference is less apparent on the larger MNIST dataset (Fig. 2c), as even a high corruption ratio leaves a sufficient absolute number of clean samples for the MLP to learn the task effectively. Therefore, while both architectures can handle this type of noise, the QNN's performance under Feature Randomization, especially in the data-scarce limit, provides further evidence of its intrinsic robustness.

https://doi.org/10.1038/s41467-026-70420-4

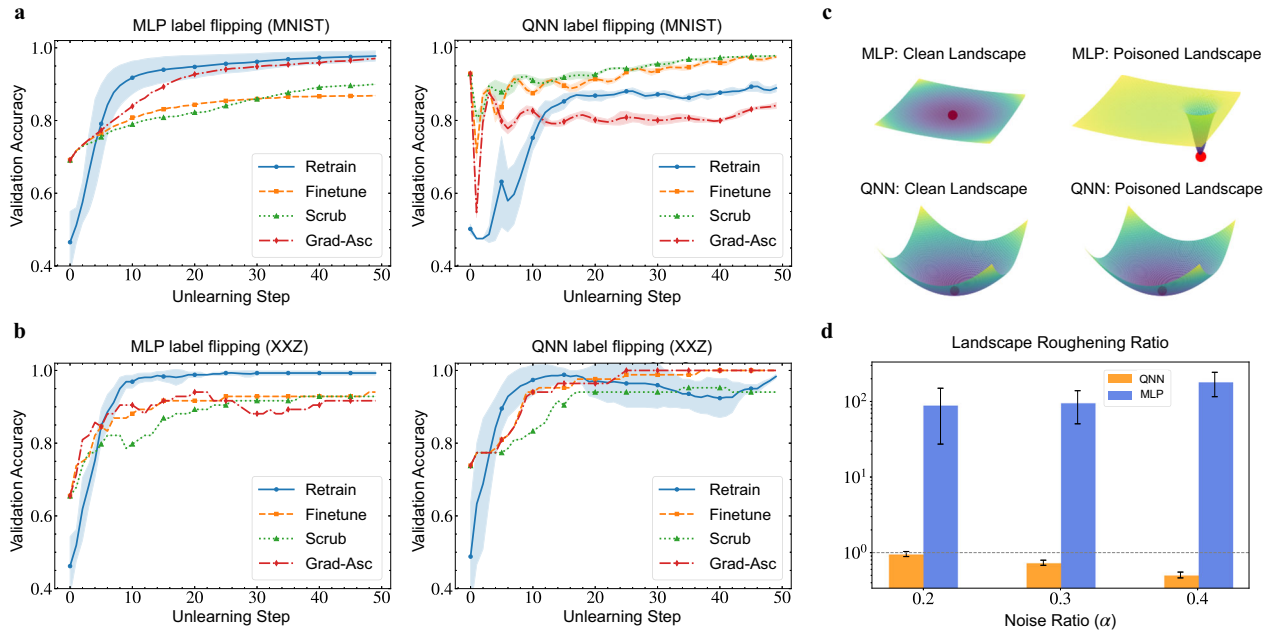

Fig. 3 | Enhanced unlearning plasticity of quantum models and its geometric origin. Unlearning dynamics for the MLP and QNN on the classical MNIST dataset (**a**) and the quantum XXZ dataset (**b**), following training on data corrupted by label flipping ($\alpha = 0.3$). MLP multi-layer perceptron, QNN quantum neural network. The plots show validation accuracy versus the unlearning step for four different unlearning methods. One unlearning step is a single training epoch over the dataset. For the MLP, the Retrain method serves as a clear upper bound, while computationally cheaper methods consistently underperform, evidencing stubborn memories. For the QNN, approximate unlearning methods are highly effective, often outperforming the costly Retrain baseline within the early training time window after a transient performance dip, demonstrating good model plasticity. Shaded regions represent the standard deviation over 5 runs. **c** Schematic illustration of the local loss landscape geometry around the local minimum of the trained model. The MLP's landscape deforms from a wide, flat minimum (Clean) into a pathologically sharp, brittle minimum (Poisoned) to memorize outliers. The QNN's robust, stable minimum remains largely unperturbed by the data corruption. **d** Quantitative analysis via the LRR, defined as the ratio of the Hessian trace in the noisy vs. clean scenarios. The classical baseline is a ReLU MLP with two hidden layers (16 and 4 units), while the QNN utilizes a 10-qubit hardware-efficient ansatz with 6 layers. The MLP's LRR is several orders of magnitude greater than one, confirming its landscape's fragility. The QNN's LRR is consistently near unity (dashed line), providing direct evidence of its landscape's structural stability. Error bars represent the standard deviation of the LRR across different initializations.

Quantum machine unlearning: plasticity versus stubborn memories

Building on the finding of the QNN's resilience, we now assess its capacity for repair by establishing the systematic investigation into quantum machine unlearning. Our unlearning protocol begins with a model initially trained on a dataset D_{train}, which is a union of a clean retain set, D_{retain}, and a polluted forget set, $D_{forget}^{polluted}$. The goal of unlearning is to efficiently alter the trained model to approximate a new model trained from scratch solely on D_{retain} with sufficient training steps, thus erasing the influence of the corrupted data. A separate, held-out clean validation set, D_{val}, is used to evaluate the final performance of the unlearned model.

To this end, we compare three efficient, approximate unlearning strategies as well as the retrain from scratch setting:

- Retrain: The baseline method that discards the polluted model entirely and trains a new one from scratch using only D_{retain}, also known as exact unlearning. It is effective but prohibitively costly in practice.
- Finetune: The simplest approximate method, which takes the polluted model and continues training on D_{retain} with the goal of overwriting the bad knowledge as implied by catastrophic forgetting[65,66].
- Scrub: A multi-objective technique involving a teacher (the original polluted model) and a student (the new model being unlearned). It simultaneously trains on D_{retain} while encouraging the student's outputs on the forget set to diverge from those of the teacher and outputs on the retain set to be consistent with those of the teacher, thus actively forgetting[67].
- Gradient Ascent: A reverse learning approach that modifies the optimization objective to perform standard gradient descent on the retain set loss while simultaneously performing gradient ascent on the forget set loss, actively unlearning the erroneous samples by reversing the training process[68].

Our unlearning studies focus on the Label Flipping scenario, as it presents the most relevant and challenging test case. In the Feature Randomization scenario, the model's performance is not significantly compromised at the low-to-moderate corruption ratios (α) typical for unlearning studies. The model effectively learns to ignore these out-of-distribution outliers, meaning the prerequisite for unlearning–a genuinely corrupted model in need of repair–is not met. Conversely, studying unlearning at very high corruption ratios ($\alpha > 0.5$) would shift the problem from "unlearning a small data subset" to "few-shot relearning from a tiny retain set", a fundamentally different paradigm. Therefore, Label Flipping at moderate α represents the most suitable testbed where the model is sufficiently damaged to necessitate repair.

The unlearning advantage of the QNN is not merely theoretical but profoundly practical. We observe a stark contrast in the unlearning dynamics between the classical and quantum models. For the classical MLP, the Retrain method quickly establishes a high-performance baseline. However, the computationally cheaper approximate unlearning methods consistently underperform, struggling to close the performance gap (Fig. 3a, b). This suggests that the MLP forms stubborn memories of the poisoned data, which cannot be easily erased without a costly hard reset via full retraining.

Conversely, for the QNN, the Retrain process from a random initialization exhibits slower convergence and greater instability. Strikingly, the approximate unlearning methods are not only more stable, but can achieve better validation accuracy within the same limited computational window than retrain from scratch (Fig. 3a, b). This demonstrates the QNN's unlearning plasticity. The pre-trained state, even though poisoned, serves as a robust starting point that is highly amenable to efficient forgetting. This highlights a more favorable trade-off between computational cost and performance recovery in the quantum model, marking a key practical advantage for quantum unlearning. We have also evaluated the accuracy on the forget set as detailed in Supplementary Information Section V.

An interesting feature of the QNN's unlearning dynamics, particularly on the MNIST dataset, is the initial, transient dip in validation performance before its subsequent recovery and ascent (Fig. 3a). We interpret this not as a flaw, but as evidence of a necessary unanchoring process. The initial poisoned state represents a stable but incorrect local minimum. The unlearning procedure acts as a targeted perturbation that temporarily increases the model's loss, effectively heating the system to allow for broader exploration of the loss landscape. This transient destabilization enables the QNN to escape the suboptimal minimum before it reconverges to a better, more correct solution. The MLP's unlearning curves lack this dynamical pattern, suggesting its stubborn memories correspond to a much deeper, inescapable minimum.

Mechanism of resilience: local loss landscape geometry

To understand the underlying mechanism behind the QNN's resilience and unlearning plasticity, we investigated the geometry of the loss landscapes found by both classical and quantum models. The flatness of a minimum in the loss landscape, which is inversely related to its sharpness, is a key indicator of good generalization and robustness to perturbations[69,70]. We characterized this geometry by computing the spectrum of the Hessian matrix ($\nabla^2\mathcal{L}$) of the loss function at the final trained model parameters. It is important to note that our analysis focuses on the local geometry around the converged minima. While the global landscapes of both classical and quantum models are non-convex and complex[71], their responses to noise around local mimima differ fundamentally.

While the absolute values of the Hessian eigenvalues are difficult to compare directly between models of such different architectures and parameter counts (around 12,600 parameters in classical MLP while only 234 parameters in QNN for MNIST task investigated here), their relative change in response to data corruption provides a clear and normalized measure of landscape stability. To quantify this, we define the Landscape Roughening Ratio (LRR) as the ratio of the Hessian matrix trace of the model trained on polluted data with that of the model trained on clean data:

$$\mathrm{LRR} = \frac{\mathrm{Tr}(\mathbf{H}_{\mathrm{noisy}})}{\mathrm{Tr}(\mathbf{H}_{\mathrm{clean}})}. \tag{1}$$

An LRR close to 1 indicates a highly robust landscape that is insensitive to noise, whereas a large LRR signifies a fragile landscape that deforms significantly toward sharpness to memorize corrupted data in a brute-force way.

Our analysis reveals a stark difference in landscape stability (Fig. 3c, d). The classical MLP exhibits extreme fragility, with a large LRR (in the order of 10^2). This astronomical increase confirms that the MLP's landscape undergoes a violent transformation to memorize the contradictory noisy labels, shifting from a very flat basin to a pathologically sharp and brittle minimum. Escaping such a sharp valley poses challenges on the unlearning algorithm. Approximate methods might fail because their gentle gradient updates are insufficient to exit this pathological basin.

In contrast, the QNN demonstrates exceptional stability. Its Hessian trace remains similar, yielding an LRR of approximately or smaller than 1. The QNN's architecture and optimization dynamics appear to resist the formation of the sharp, complex minima required to overfit to outliers. It preserves the clean landscape geometry, thereby maintaining its robust generalization performance. This inherent structural stability, as captured by its near-unity LRR, is the origin of both its resilience to poisoning and its amenability to efficient unlearning. A further open question arises when considering the scaling of these models. While deeper circuits may offer greater expressivity, they also risk the onset of barren plateaus–regions of extreme flatness[72–77]. Future research must therefore explore a critical trade-off: does the flatness inherent to barren plateaus represent an ultimate form of robustness, or does it constitute a trivial stability that comes at the cost of all learnable features?

The empirical LRR findings are further supported by a rigorous analytical derivation (see Supplementary Information Section I). We prove that deep classical ReLU networks are susceptible to an unbounded curvature explosion, allowing them to form arbitrarily sharp minima to memorize outliers. In contrast, QNNs possess structural stiffness where the Hessian is bounded by the unitary nature of quantum gates. Furthermore, we derive an error-phase interaction term which generates negative definite curvature when the model attempts to fit contradictory labels, actively canceling out the positive curvature required for memorization. This mechanism mathematically prevents the formation of sharp, brittle minima for data noise, forcing the QNN to settle for a smooth, generalizable solution.

Discussion

Our findings show intrinsic differences between two learning paradigms. In the face of corrupted data, the classical MLP acts as a diligent but brittle stenographer, meticulously recording every detail, including falsehoods. The QNN, in contrast, behaves more like a discerning editor, prioritizing the integrity of the main narrative over accommodating every outlier. Our analysis of the loss landscape geometry provides a compelling mechanism for this difference, revealing that the QNN possesses a dual advantage rooted in a single, fundamental property: the inherent structural stability of its landscape.

Our central finding that QNNs exhibit a resilience to data corruption can be understood as the manifestation of the strong generalization properties for QML models[64]. The theoretical guarantee of a small generalization error is not merely a passive property for clean data; our results show that it is an active one. The QNN fights to preserve its generalizable solution even when pushed towards memorization by label noise. Our Hessian analysis reveals the physical mechanism that underpins this robust generalization. While the concept of flat minima is well-established for generalization, we introduce the crucial dimension of landscape stability under data corruption. The classical MLP finds an exceedingly flat but fragile minimum on clean data, which deforms into a pathologically sharp landscape under data noise. In contrast, the QNN's loss landscape is structurally stable, with its geometry remaining almost entirely unperturbed by data corruption, as quantified by its near-unity LRR. This stability is the direct cause of the favorable generalization guarantees for QML. The macroscopic phase transition we observe at $\alpha = 0.5$ is the ultimate signal of this stability: the system resists disorder up to a critical point before undergoing a cooperative shift, a behavior far more structured and robust than the continuous performance compromise of the MLP.

Our finding that QNNs exhibit a robust resilience to label noise, preferring to maintain a simple, generalizable solution rather than memorizing outliers, may seem to be in tension with recent work demonstrating that QNNs possess a high capacity for memorization[78], where QNN is shown to be expressive enough to fit pure noise. In fact, our work provides crucial complementary insight into the model's inductive bias. While Ref. 78 establishes what QNNs can do (capacity), our comparative study with MLPs reveals what they prefer to do. Our

results show that when a dataset contains a mixture of signal and noise, QNN creates a strong preference for the signal. The MLP, on the other hand, can also remember all training samples[79] while lacking this robust inductive bias and readily engages in brittle memorization of the noise. Therefore, the QML generalization is twofold: QNNs possess a surprisingly high capacity to memorize, yet they simultaneously exhibit a built-in resistance to doing so when a simpler, more general solution exists. The high potential capacity coupled with a strong regularization bias is precisely what makes QNN promising for learning in noisy, real-world environments.

These results compel us to reconsider the nature of the quantum advantage in the near term. While the quest for exponential speedups remains an ultimate goal, our work suggests that a more immediate and practical advantage may lie in the domains of robustness and trustworthiness. In an information ecosystem rife with noise, bias, and adversarial manipulations, the ability of a model to resist corruption and be efficiently corrected is of great importance. The dual advantage of resilience and plasticity positions QML as a uniquely promising paradigm for building reliable AI systems.

However, our study carves out only small pieces of this vast territory. We operated under the assumption of fully known forget sets and noise at the data level. This opens several crucial future directions. A primary question is to disentangle the effects of data-level noise from device-level quantum noise. How does the inherent noise of near-term hardware interact with our observed resilience? Does quantum noise act as a further regularizer, enhancing robustness, or does it degrade the model's ability to distinguish signal from noise, potentially erasing this advantage? Furthermore, future work should also explore these phenomena across a wider range of model architectures, datasets, learning tasks, and unlearning techniques, specifically for unsupervised learning and generative tasks. By continuing to chart the response of quantum models to real-world data imperfections, we can pave the way for secure and adaptable quantum technologies.

Methods

Classification tasks

In this study, we investigate two distinct binary classification problems to evaluate the performance of our methods.

XXZ model ground state classification. This problem originates from quantum many-body physics and involves classifying the quantum ground states of the one-dimensional spin-1/2 Heisenberg XXZ model. The system consists of a chain of $L = 12$ spins with periodic boundary conditions. The Hamiltonian of the system is given by:

$$H_{XXZ} = \sum_{i=1}^{L} \left(\sigma_i^x \sigma_{i+1}^x + \sigma_i^y \sigma_{i+1}^y + \Delta \sigma_i^z \sigma_{i+1}^z\right), \tag{2}$$

where σ_i^α are the Pauli matrices for the spin at site i, and Δ is the anisotropy parameter that controls the phase of matter. The boundary condition imposes that $\sigma_{L+1} \equiv \sigma_1$.

The dataset is generated by numerically calculating the ground state eigenvector of this Hamiltonian for different values of Δ using exact diagonalization. The ground state, a normalized vector in $\mathbb{C}^{2^L}$, serves as the input feature vector $\mathbf{x}$. The task is a binary classification problem to distinguish between two quantum phases based on the anisotropy Δ:

- Class 0 (Gapless Phase): The ground states generated for values of the anisotropy parameter Δ in the range:

$$\Delta \in [-0.96, -0.94, \ldots, 0.94, 0.96].$$

  This corresponds to the gapless (critical) phase of the XXZ model.

- Class 1 (Gapped Phase): The ground states generated for values of the anisotropy parameter Δ in the range:

$$\Delta \in [1.02, 1.04, \ldots, 2.98, 3.00].$$

  This corresponds to the gapped antiferromagnetic Ising-like phase.

The validation dataset includes the ground states from $\Delta \in [-0.97, -0.95, \ldots, 2.99]$.

MNIST digit classification. This is a classical computer vision task based on a subset of the well-known MNIST dataset of 28 × 28 handwritten digits.

The task is a binary classification problem designed to distinguish between two specific digits, '1' and '9'. To create a balanced dataset, we select 250 image samples for each of these two digits. The input feature vector is the flattened pixel representation of the image. The classes are defined as follows:

- Class 0: 250 images of the digit '1'.
- Class 1: 250 images of the digit '9'.

The validation dataset includes 1000 other images of the two digits.

Model architectures

We compare the classical MLP with QNN. The specific hyperparameters for each model were chosen to achieve effective learning on their respective tasks (see Supplementary Information Section VII). The schematic architectures are shown in Fig. 4.

MLP. Our classical model is a standard feed-forward multi-layer perceptron, used for both the XXZ and MNIST tasks. The architecture consists of an input layer accepting the flattened data vectors, followed by two hidden layers. The Rectified Linear Unit (ReLU) serves as the activation function for both hidden layers. The output layer consists of a single neuron with a sigmoid activation function to produce a probability for the binary classification task. For the XXZ model ground state classification task, the wavefunction components are split into real and imaginary parts for the pure real input of MLP.

QNN. Our quantum model is a parameterized quantum circuit, implemented using the TensorCircuit-NG library[80]. The architecture of the PQC was adapted to the specific dataset to ensure effective learning.

- For the XXZ task, the model operates on a register of 12 qubits, matching the spin chain length. The 4096-dimensional ground state wavefunction is loaded directly into the initial state via amplitude encoding. The variational ansatz has a depth of 4.
- For the MNIST task, the model operates on 10 qubits. The 1024-dimensional padded input vector is loaded via amplitude encoding. To handle the increased complexity of this classical dataset, the variational ansatz has a depth of 6.

For both tasks, each layer of the ansatz is composed of single-qubit rotations (RX, RY, and RX gates) applied to all qubits, followed by a layer of RZZ gates ($e^{i\frac{\theta}{2}Z_jZ_{j+1}}$) arranged in a linear chain to generate entanglement. The classification output is derived from the expectation value of the Pauli-Z operator measured on the final qubit, which is then passed through a scaled sigmoid function to map the output to a probability.

Data contamination

To simulate real-world scenarios where training data may contain errors, we employ two distinct methods to contaminate an originally clean dataset.

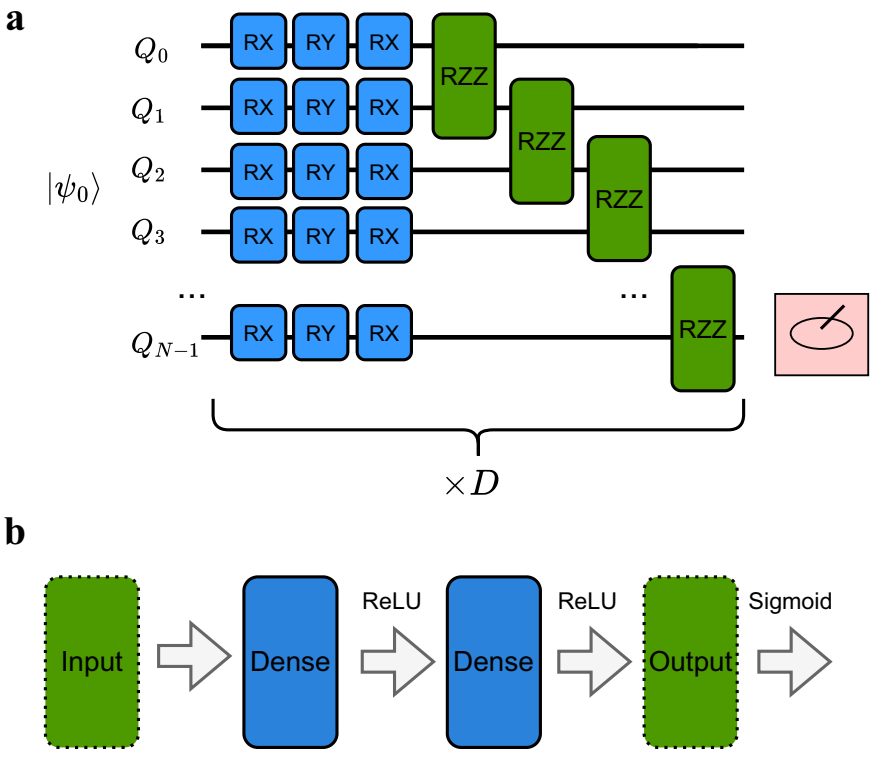

Fig. 4 | Model architectures. a Schematic of the QNN architecture used in this study. The model consists of an N-qubit register initialized with an input state $|\psi_0\rangle$. A variational ansatz is applied, composed of D repeated layers. Each layer consists of single-qubit rotations (RX, RY, RX) followed by entangling two-qubit gates (RZZ) applied to adjacent qubits. A final measurement is performed to extract the classical output. **b** Diagram of the MLP architecture. The classical model is a feed-forward neural network consisting of an input layer, two fully-connected (Dense) hidden layers each followed by a ReLU activation, and a final output layer with a sigmoid activation function for binary classification.

Label flipping. This method targets the labels (y) of the data, simulating common annotation errors. For a binary classification task where the labels are $y \in \{0, 1\}$, the label flipping operation is defined as:

$$y_{\text{poisoned}} = 1 - y_{\text{clean}}. \tag{3}$$

This directly corrupts the mapping between features and labels, rendering the model to learn an incorrect, and potentially opposite, decision boundary. It represents a challenging form of data poisoning.

Feature randomization. This method targets the features ($\mathbf{x}$) of the data, simulating corrupted or replaced input. For a subset of samples of ratio α, the original feature vector $\mathbf{x} \in \mathbb{C}^d$ is replaced by a randomly generated vector, $\mathbf{x}'_{\text{poison}}$. The random vector is real for MNIST task and complex for XXZ model ground state task. For the complex one, the generation process is as follows:

1. Two real-valued vectors, $\boldsymbol{\theta}_{\text{real}}$ and $\boldsymbol{\theta}_{\text{imag}}$, are drawn from a normal (Gaussian) distribution, where each component is sampled independently:

$$(\boldsymbol{\theta}_{\text{real}})_j, (\boldsymbol{\theta}_{\text{imag}})_j \sim \mathcal{N}(\mu = 0, \sigma^2 = 1). \tag{4}$$

2. The poisoned input vector is then constructed using these random vectors as its real and imaginary parts and finally get normalized:

$$\mathbf{x}'_{\text{poison}} = \boldsymbol{\theta}_{\text{real}} + i \cdot \boldsymbol{\theta}_{\text{imag}}. \tag{5}$$

This method completely destroys the original feature structure, forcing the model to associate a label with meaningless random noise, which can severely damage the learned feature representations.

In the context of our machine unlearning experiments, the data is partitioned into the following sets:

- $\mathcal{D}_{\textbf{train}}$ (Training Set): This is the dataset used to train the initial, polluted model. It is a composite set containing both the clean retain data and the polluted forget data. Formally, $\mathcal{D}_{\text{train}} = \mathcal{D}_{\text{retain}} \cup \mathcal{D}^{\text{polluted}}_{\text{forget}}$.
- $\mathcal{D}_{\textbf{retain}}$ (Retain Set): This is the subset of the original training data whose knowledge the model should retain after the unlearning process. This set is entirely clean and unpolluted. A primary goal of any unlearning algorithm is to maintain high performance on $\mathcal{D}_{\text{retain}}$.
- $\mathcal{D}_{\textbf{forget}}$ (Forget Set): This is the subset of the original training data whose influence the model must forget. During the initial training, the polluted version of this set, $\mathcal{D}^{\text{polluted}}_{\text{forget}}$, is used. For evaluation after unlearning, the corresponding clean version, $\mathcal{D}^{\text{clean}}_{\text{forget}}$, is typically used to verify that the model can now correctly classify this type of data. It is important to note that for unlearning methods like Gradient Ascent, the polluted version, $\mathcal{D}^{\text{polluted}}_{\text{forget}}$, is used during the unlearning phase to directly counteract the original erroneous learning signals.
- $\mathcal{D}_{\textbf{val}}$ (Validation Set): This is a completely clean dataset that is held out from all training and unlearning procedures. It serves as the gold standard for evaluating the final performance of the unlearned model, as it measures the model's ability to generalize to unseen clean data after the unlearning process is complete.

Machine unlearning methods

To assess the capacity of each model for repair, we implement and compare four distinct machine unlearning algorithms. Each procedure begins with a model pre-trained on a corrupted dataset and aims to remove the influence of a specified forget set, $\mathcal{D}_{\text{forget}}$, while preserving

knowledge from the clean retain set, $\mathcal{D}_{\text{retain}}$. We define one unlearning step as a single training epoch over the retain set $\mathcal{D}_{\text{retain}}$ (and the forget set $\mathcal{D}_{\text{forget}}$, for methods that utilize it).

Retrain. Considered the gold standard for unlearning, this method involves completely discarding the polluted model and training a new model from a random initialization, using only the clean retain set $\mathcal{D}_{\text{retain}}$. Its objective is to minimize the standard empirical risk on this sanitized data:

$$\boldsymbol{\theta}_{\text{new}} = \arg\min_{\boldsymbol{\theta}} \sum_{(\mathbf{x},y)\in\mathcal{D}_{\text{retain}}} \mathcal{L}(\boldsymbol{\theta};\mathbf{x},y). \tag{6}$$

While providing a perfect removal of the forget set's influence, this approach is often computationally prohibitive for large-scale models.

Finetune. This is a computationally efficient approximation of retraining. The procedure starts with the parameters of the already-trained polluted model, θ_{old}, and continues the training process using only the clean retain set $\mathcal{D}_{\text{retain}}$. The objective is to overwrite the bad knowledge learned from the forget set by leveraging the phenomenon of catastrophic forgetting. The loss function is identical to that of the Retrain method, but the optimization starts from $\boldsymbol{\theta}_{\text{old}}$ instead of a random initialization.

Gradient ascent. This method implements a more targeted reverse learning process. It unifies the objectives of retaining and forgetting into a single loss function. The model simultaneously performs standard gradient descent on the retain set while performing gradient ascent on the forget set. This is achieved by minimizing a composite loss where the contribution from the forget set is subtracted:

$$\mathcal{L}_{GA}(\theta) = \sum_{(\mathbf{x},y)\in\mathcal{D}_{\text{retain}}} \mathcal{L}(\boldsymbol{\theta};\mathbf{x},y) - \beta \sum_{(\mathbf{x},y)\in\mathcal{D}_{\text{forget}}^{\text{polluted}}} \mathcal{L}(\boldsymbol{\theta};\mathbf{x},y). \tag{7}$$

Here, β is a hyperparameter that balances the strength of forgetting against retaining. This unified objective often leads to a more stable optimization process than multi-objective techniques.

Scrub. This technique employs a multi-objective push-pull framework that explicitly separates the goals of retaining and forgetting[67]. It uses the original polluted model, $\boldsymbol{\theta}_{\text{old}}$, as a teacher to guide the unlearning of a student model, $\boldsymbol{\theta}$. The total loss is a weighted sum of three distinct components:

$$\begin{aligned}\mathcal{L}_{\text{scrub}}(\boldsymbol{\theta}) \quad = \quad & \lambda_{ce}\mathcal{L}_{CE}(\boldsymbol{\theta};\mathcal{D}_{\text{retain}}) \\ & +\lambda_{kl}\mathcal{L}_{KL,\text{retain}}(\boldsymbol{\theta}) \\ & -\lambda_{fo}\mathcal{L}_{KL,\text{forget}}(\boldsymbol{\theta}).\end{aligned} \tag{8}$$

The first term is a standard cross-entropy loss on the retain set to maintain accuracy. The second term, $\mathcal{L}_{KL,\text{retain}}$, uses the Kullback-Leibler (KL) divergence to encourage the student's outputs on the retain set to remain close to the teacher's, thereby anchoring useful knowledge. The final term, $\mathcal{L}_{KL,\text{forget}}$, maximizes the KL divergence between the student's and teacher's outputs on the forget set, explicitly pushing the student model away from the teacher's incorrect predictions. The hyperparameters λ_{ce}, λ_{kl}, and λ_{fo} must be carefully tuned to balance these often-conflicting objectives.

We have evaluated these unlearning methods with different hyperparameters and multiple trials, please see Supplementary Information Section VI for details. All numerical results are carried out on high performance computing clusters.

Data availability

The data generated in this study have been deposited in the Zenodo database under accession link https://doi.org/10.5281/zenodo.18641024[81]. The MNIST dataset is publicly available at http://yann.lecun.com/exdb/mnist/.

Code availability

The custom code and scripts used to generate the results and analyze the data in this study are publicly available on GitHub at https://github.com/yutuer21/quantum-machine-unlearningand have been permanently archived on Zenodo under the https://doi.org/10.5281/zenodo.18641024[81]. The quantum neural network implementations are based on the open-source quantum software framework TensorCircuit-NG, which is available at https://github.com/tensorcircuit/tensorcircuit-ng.

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436 (2015).
2. Jordan, M. I. & Mitchell, T. M. Machine learning: trends, perspectives, and prospects. *Science* **349**, 255 (2015).
3. Vaswani, A. et al. Attention is all you need. In *NeurIPS* (NIPS, 2017).
4. Barreno, M., Nelson, B., Sears, R., Joseph, A. D. & Tygar, J. D. Can machine learning be secure? In *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security*, 16 (ACM, 2006).
5. Goldblum, M. et al. Dataset security for machine learning: data poisoning, backdoor attacks, and defenses. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 1563 (2022).
6. Schwarzschild, A., Goldblum, M., Gupta, A., Dickerson, J. P. & Goldstein, T. Just how toxic is data poisoning? A unified benchmark for backdoor and data poisoning attacks. In *Proc. 38 th International Conference on Machine Learning* (PMLR, 2021).
7. Tian, Z., Cui, L., Liang, J. & Yu, S. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Comput. Surv.* **55**, 1 (2023).
8. Zhao, P., Zhu, W., Jiao, P., Gao, D. & Wu, O. Data poisoning in deep learning: a survey. Preprint at https://arxiv.org/abs/2503.22759 (2025).
9. Cao, Y. & Yang, J. in *2015 IEEE Symposium on Security and Privacy*, 463–480 (IEEE, 2015).
10. Ginart, A., Guan, M. Y., Valiant, G. & Zou, J. Making AI forget you: data deletion in machine learning. In *NeurIPS* (NIPS, 2019).
11. Bourtoule, L. et al. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)* 141 (IEEE, 2021).
12. Wang, W., Tian, Z., Zhang, C. & Yu, S. Machine unlearning: a comprehensive survey. Preprint at https://arxiv.org/abs/2405.07406 (2024).
13. Goel, S., Prabhu, A., Torr, P., Kumaraguru, P. & Sanyal, A. Corrective machine unlearning. Preprint at https://arxiv.org/abs/2402.14015 (2024).
14. Liu, S. et al. Rethinking machine unlearning for large language models. *Nat. Mach. Intell.* **7**, 181 (2025).
15. Qu, Y. et al. The frontier of data erasure: a survey on machine unlearning for large language models. *Computer* **58**, 45 (2025).
16. Preskill, J. Quantum computing in the nisq era and beyond. *Quantum* **2**, 79 (2018).
17. Bharti, K. et al. Noisy intermediate-scale quantum algorithms. *Rev. Mod. Phys.* **94**, 015004 (2022).
18. Cerezo, M. et al. Variational quantum algorithms. *Nat. Rev. Phys.* **3**, 625 (2021).
19. Biamonte, J. et al. Quantum machine learning. *Nature* **549**, 195 (2017).
20. Lloyd, S., Mohseni, M. & Rebentrost, P. Quantum algorithms for supervised and unsupervised machine learning. Preprint at https://arxiv.org/abs/1307.0411 (2013).

21. Cai, X.-D. et al. Entanglement-based machine learning on a quantum computer. *Phys. Rev. Lett.* **114**, 110504 (2015).
22. Havlíček, V. et al. Supervised learning with quantum-enhanced feature spaces. *Nature* **567**, 209 (2019).
23. Gao, X., Zhang, Z.-Y. & Duan, L.-M. A quantum machine learning algorithm based on generative models. *Sci. Adv.* **4**, aat9004 (2018).
24. Li, J., Yang, X., Peng, X. & Sun, C.-P. Hybrid quantum-classical approach to quantum optimal control. *Phys. Rev. Lett.* **118**, 150503 (2017).
25. Huang, H.-Y. et al. Power of data in quantum machine learning. *Nat. Commun.* **12**, 2631 (2021).
26. Huang, H.-Y. et al. Quantum advantage in learning from experiments. *Science* **376**, 1182 (2022).
27. Zhang, S.-X. et al. Variational quantum-neural hybrid eigensolver. *Phys. Rev. Lett.* **128**, 120502 (2022).
28. Chen, Y.-Q., Chen, Y., Lee, C.-K., Zhang, S. & Hsieh, C.-Y. Optimizing quantum annealing schedules with monte carlo tree search enhanced with neural networks. *Nat. Mach. Intell.* **4**, 269 (2022).
29. Cerezo, M., Verdon, G., Huang, H.-Y., Cincio, L. & Coles, P. J. Challenges and opportunities in quantum machine learning. *Nat. Comput. Sci.* **2**, 567 (2022).
30. Wang, Y. & Liu, J. A comprehensive review of quantum machine learning: from nisq to fault tolerance. *Rep. Prog. Phys.* **87**, 116402 (2024).
31. Schuld, M., Sinayskiy, I. & Petruccione, F. The quest for a quantum neural network. *Quantum Inf. Process.* **13**, 2567 (2014).
32. Romero, J., Olson, J. P. & Aspuru-Guzik, A. Quantum autoencoders for efficient compression of quantum data. *Quantum Sci. Technol.* **2**, 045001 (2017).
33. Liu, J.-G. & Wang, L. Differentiable learning of quantum circuit born machines. *Phys. Rev. A* **98**, 062324 (2018).
34. Benedetti, M., Lloyd, E., Sack, S. & Fiorentini, M. Parameterized quantum circuits as machine learning models. *Quantum Sci. Technol.* **4**, 043001 (2019).
35. Beer, K. et al. Training deep quantum neural networks. *Nat. Commun.* **11**, 808 (2020).
36. Shen, H., Zhang, P., You, Y.-Z. & Zhai, H. Information scrambling in quantum neural networks. *Phys. Rev. Lett.* **124**, 200504 (2020).
37. Dallaire-Demers, P.-L. & Killoran, N. Quantum generative adversarial networks. *Phys. Rev. A* **98**, 012324 (2018).
38. Cong, I., Choi, S. & Lukin, M. D. Quantum convolutional neural networks. *Nat. Phys.* **15**, 1273 (2019).
39. Pérez-Salinas, A., Cervera-Lierta, A., Gil-Fuster, E. & Latorre, J. I. Data re-uploading for a universal quantum classifier. *Quantum* **4**, 226 (2020).
40. Banchi, L., Pereira, J. & Pirandola, S. Generalization in quantum machine learning: a quantum information standpoint. *PRX Quantum* **2**, 040321 (2021).
41. Zhang, S.-X., Hsieh, C.-Y., Zhang, S. & Yao, H. Neural predictor based quantum architecture search. *Mach. Learn. Sci. Technol.* **2**, 045027 (2021a).
42. Li, W. & Deng, D.-L. Recent advances for quantum classifiers. *Sci. China Phys. Mech. Astron.* **65**, 220301 (2022).
43. Zheng, P.-L., Wang, J.-B. & Zhang, Y. Efficient and quantum-adaptive machine learning with fermion neural networks. *Phys. Rev. Appl.* **20**, 044002 (2023).
44. Jäger, J. & Krems, R. V. Universal expressiveness of variational quantum classifiers and quantum kernels for support vector machines. *Nat. Commun.* **14**, 576 (2023).
45. Miao, J., Hsieh, C.-Y. & Zhang, S.-X. Neural-network-encoded variational quantum algorithms. *Phys. Rev. Appl.* **21**, 014053 (2024).
46. Zhang, B., Xu, P., Chen, X. & Zhuang, Q. Generative quantum machine learning via denoising diffusion probabilistic models. *Phys. Rev. Lett.* **132**, 100602 (2024a).
47. Sein, P. S. S., Cañizo, M. & Orús, R. Image classification with rotation-invariant variational quantum circuits. *Phys. Rev. Res.* **7**, 013082 (2025).
48. Zhang, P. Correcting a noisy quantum computer using a quantum computer. Preprint at https://arxiv.org/abs/2506.08331 (2025).
49. Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I. & Tygar, J. D. Adversarial machine learning. In *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, 43 (ACM, 2011).
50. Biggio, B. et al. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2013. Lecture Notes in Computer Science*. (eds Blockeel, H., Kersting, K., Nijssen, S. & Železný, F.) Vol. 8190, 387–402 (Springer, Berlin, Heidelberg, 2013).
51. Szegedy, C. et al. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations* (ICLR, 2014).
52. Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and Harnessing Adversarial Examples. In *3rd International Conference on Learning Representations* (ICLR, 2015).
53. Kurakin, A., Brain, G., J, G. I. & Bengio, S. Adversarial machine learning at scale. In *ICLR* (ICLR, 2017).
54. Lu, S., Duan, L.-M. & Deng, D.-L. Quantum adversarial machine learning. *Phys. Rev. Res.* **2**, 033212 (2020).
55. Liu, N. & Wittek, P. Vulnerability of quantum classification to adversarial perturbations. *Phys. Rev. A* **101**, 062331 (2020).
56. Du, Y., Hsieh, M.-H., Liu, T., Tao, D. & Liu, N. Quantum noise protects quantum classifiers against adversaries. *Phys. Rev. Res.* **3**, 023153 (2021).
57. Weber, M., Liu, N., Li, B., Zhang, C. & Zhao, Z. Optimal provable robustness of quantum classification via quantum hypothesis testing. *npj Quantum Inf.* **7**, 76 (2021).
58. Liao, H., Convy, I., Huggins, W. J. & Whaley, K. B. Robust in practice: adversarial attacks on quantum machine learning. *Phys. Rev. A* **103**, 042427 (2021).
59. Ren, W. et al. Experimental quantum adversarial learning with programmable superconducting qubits. *Nat. Comput. Sci.* **2**, 711 (2022).
60. Gong, W. & Deng, D.-L. Universal adversarial examples and perturbations for quantum classifiers. *Natl. Sci. Rev.* **9**, nwab130 (2021).
61. West, M. T. et al. Benchmarking adversarially robust quantum machine learning at scale. *Phys. Rev. Res.* **5**, 023186 (2023a).
62. West, M. T. et al. Towards quantum enhanced adversarial robustness in machine learning. *Nat. Mach. Intell.* **5**, 581 (2023b).
63. Yang, J., Zhou, K., Li, Y. & Liu, Z. Generalized out-of-distribution detection: a survey. *Int. J. Comput. Vis.* **132**, 5635 (2024).
64. Caro, M. C. et al. Generalization in quantum machine learning from few training data. *Nat. Commun.* **13**, 4919 (2022).
65. French, R. Catastrophic forgetting in connectionist networks. *Trends Cogn. Sci.* **3**, 128 (1999).
66. Goel, S. et al. Towards adversarial evaluations for inexact machine unlearning. Preprint at https://arxiv.org/abs/2201.06640 (2023).
67. Kurmanji, M., Triantafillou, P., Hayes, J. & Triantafillou, E. *Towards Unbounded Machine Unlearning* (NeuIPS, 2023).
68. Trippa, D., Campagnano, C., Bucarelli, M. S., Tolomei, G. & Silvestri, F. $\nabla\tau$: Gradient-based and task-agnostic machine unlearning. Preprint at https://arxiv.org/abs/2403.14339 (2024).
69. Hochreiter, S. & Schmidhuber, J. Flat minima. *Neural Comput.* **9**, 1 (1997).
70. Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M. & Tang, P. T. P. On large-batch training for deep learning: generalization gap and sharp minima. In *ICLR* (ICLR, 2017).
71. Li, H., Xu, Z., Taylor, G., Studer, C. & Goldstein, T. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18*, 6391–6401 (Curran Associates Inc., 2018).

72. McClean, J. R., Boixo, S., Smelyanskiy, V. N., Babbush, R. & Neven, H. Barren plateaus in quantum neural network training landscapes. *Nat. Commun.* **9**, 4812 (2018).
73. Wang, S. et al. Noise-induced barren plateaus in variational quantum algorithms. *Nat. Commun.* **12**, 6961 (2021).
74. Cerezo, M. & Coles, P. J. Higher order derivatives of quantum neural networks with barren plateaus. *Quantum Sci. Technol.* **6**, 035006 (2021).
75. Zhang, H.-K., Liu, S. & Zhang, S.-X. Absence of barren plateaus in finite local-depth circuits with long-range entanglement. *Phys. Rev. Lett.* **132**, 150603 (2024b).
76. Zhang, H.-K., Zhu, C. & Wang, X. Predicting quantum learnability from landscape fluctuation. Preprint at https://arxiv.org/abs/2406.11805 (2024c).
77. Larocca, M. et al. Barren plateaus in variational quantum computing. *Nat. Rev. Phys.* **7**, 174 (2025).
78. Gil-Fuster, E., Eisert, J. & Bravo-Prieto, C. Understanding quantum machine learning also requires rethinking generalization. *Nat. Commun.* **15**, 2277 (2024).
79. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM* **64**, 107 (2021b).
80. Zhang, S.-X. et al. Tensorcircuit: a quantum software framework for the nisq era. *Quantum* **7**, 912 (2023).
81. Chen, Y.-Q. & Zhang, S.-X. Superior resilience to poisoning and amenability to unlearning in quantum machine learning. Zenodo repository, https://doi.org/10.5281/zenodo.18641024 (2026).

Acknowledgements

This work was supported by Science Challenge Project (No.TZ2025017 [Y.Q.C.]), Quantum Science and Technology-National Science and Technology Major Project (No. 2024ZD0301700 [S.X.Z.]), the National Natural Science Foundation of China (Nos. 12504599 [Y.Q.C.] and 12574546 [S.X.Z.]), and NSAF (No. U2330401 [Y.Q.C.]).

Author contributions

Y.Q.C. and S.X.Z. conceived the project, designed and performed the research, analyzed the data, and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-70420-4.

Correspondence and requests for materials should be addressed to Yu-Qin Chen or Shi-Xin Zhang.

Peer review information : *Nature Communications* thanks Tongyang Li and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.