

---

## Reviewer 1

**Comment.** *The paper is concerned with Zero-noise extrapolation (ZNE) for addressing the challenge of gate noise in quantum circuits. ZNE increases quantum circuit noise and extrapolates the measurement outcomes to the zero-noise limit, which may be costly. The paper introduces a new student–teacher knowledge distillation approach for error prevention and compression. This conceptually new method is analyzed mathematically and experimentally using both simulated noise and real quantum hardware.*

### Reviewer Point P 1.1 — Mismatch between provided source code and details in the paper.

- The source code is using qiskit version 0.45.3 (according to its “requirements.txt” file), which is depreciated since February 2024.
- The code is not consistent at all with the procedure described in Section 2.2.1. For example, consider the loss described in Step (iii) ”Distillation Objectives” using tanh and ZNE corrected expectations. In contrast, in the code (take wine/kd.py) in sections 1.6.-1.8 a bigger QNN is trained, then used to generate predictions, and then a smaller QNN is trained on those predictions.
- The code also does not include any of the benchmarks reported in Table 4. On the other hand, some of these benchmarks (e.g. Wang et al. 2025) do not include experiments with the datasets considered here. This heavily undermines the paper’s claims.

**Response:** We appreciate the reviewer’s observations.

- The original code used `qiskit==0.45.3`, which was deprecated in February 2024. However, it is still fully functional and interfaces correctly with both `qiskit-aer` and the `qiskit-ibm-runtime` service needed for real experiments on `IBMBrisbane`. This version was chosen to provide stable compatibility with Aer simulations and IBM Runtime hardware APIs during the evaluations.

`qiskit = 1.0.2, qiskit-aer = 0.13.1, and qiskit-ibm-runtime = 0.22.0`  
packages are the updated version and not compatible with each-other.

- We acknowledge that the simplified demo code used standard KD—training a larger teacher QNN, generating predictions, and training a smaller student on those predictions—whereas Section 2.2.1 described the full ZNKD formulation, including the tanh-stabilized regression loss and ZNE-corrected soft targets. To remove the discrepancy, we changed Section 2.2.1 so that the wording matches the methodology utilized in the code.

The tanh initially developed because student QNNs with smaller widths and fewer parameters frequently had a shorter output range and higher gradient saturation. Using tanh, teacher

outputs are transformed into a smooth, bounded interval  $(-1, 1)$  that corresponds to the representational range of shallow variational circuits and displays the ideal objective function.

However, in code, we simply implemented a very lightweight version of the Quantum model. We have updated the implementation details to avoid confusion as follows.

**iii) Distillation objectives.** Our distillation approach uses the conventional two-stage quantum knowledge distillation process. To create a high-expressivity model, a larger VQC teacher is taught using ground-truth labels. Following convergence, the teacher makes predictions for all training samples as

$$\hat{y}_T(x) = \text{Teacher}(x).$$

*Hard-label distillation.* The teacher’s arg-max predictions are used to train the student VQC in the default configuration as

$$\hat{y}_S(x) \approx \arg \max_b \hat{y}_T^b(x),$$

which is equivalent to reducing the mean-squared error (MSE) between the teacher’s discrete label assignments and the student’s output.

- Thank you for bringing this to our attention. Table 4 had an unintentional citation error involving Wang et al. (2025). Li et al. (2024) should be cited in the appropriate benchmark comparison. We have updated the table.

Methods	Ref.	F-MNIST ↓	AG News ↓	Wine Quality ↓	Urban Sound8K ↓
Noise-aware VQCs	[1]	0.56	0.49	0.59	0.65
Pruned QNNs	[2]	0.53	0.46	0.55	0.62
Classical-to-Q KD	[3]	0.51	0.40	0.53	0.61
ZNE Only	[4]	0.52	0.42	0.54	0.60
Best-Folding ZNE	[5]	0.50	0.38	0.52	0.59
PEC	[6]	0.50	0.39	0.53	0.59
MEM	[7]	0.51	0.38	0.55	0.62
CDR	[8]	0.52	0.39	0.54	0.60
DD	[9]	0.55	0.42	0.57	0.70
Hardware-Efficient Ansätze	[10]	0.54	0.41	0.56	0.65
<b>Ours (ZNKD)</b>	–	<b>0.49</b>	<b>0.36</b>	<b>0.52</b>	<b>0.59</b>

Table 1: Comparison with state-of-the-art quantum noise mitigation and compression approaches under identical conditions. Our full method (ZNKD) achieves the lowest loss across all datasets, demonstrating the synergy between circuit-level extrapolation and quantum-native distillation.

## Reviewer Point P 1.2 — Style and presentation issues.

- Writing and presentation: The level of presentation is currently below ICLR standards. Section 2 requires the reader to jump back and forth between Supplement and Main text several

times within the first lines of reading. The main paper should be self-contained, with supplement providing additional background and further supporting material. However, the supplement should not be needed for a basic understanding of what the paper aims to do. For example, the presentation of the method in Section 2.1 starts with "After defining gate-level decoherence using the Lindblad-informed noise model in A.1 ..." Then in "Motivation" it goes on "Using the single-gate fidelity equation 32, demonstrates...". The same happens in several places below.

- Typos: The paper still contains many typos, such as typesetting of "U" in l102 and expressions like "These denoised outputs are used as soft labels for training students (clients)." (what does "clients" refer to here?). Section 3.2 (lines 403-405) contain several misplaced "!"-symbols.
- References out of place: In line 90-91 ZNE is attributed to a paper from 2024. However, this does not appear to be the correct reference. Another example is Wang et al. (2025), which (differently than stated in the text) is not concerned with knowledge distillation or quantum transfer learning. Nevertheless, it is listed as a benchmark method in Table 4.
- The paper seems to still contain an LLM prompt: line 284-285 states: "Give an explanation of the extrapolated estimator's mean squared error (MSE) as ...".

#### Response:

- Thank you for your valuable observation. We agree that Section 2's dependence on Supplementary material (A.1 and Equations 30 and 32) was overbearing, requiring readers to go back and forth in order to comprehend the fundamental approach. In order to address this, we have revised Section 2.1 to explicitly include all necessary concepts in the main study, including the noise model, gate-level decoherence, circuit-level error behavior, and the justification for ZNE. Only expanded derivations are now included in the Supplement. This ensures that the main document can be read without consulting the appendix and is self-contained.

After defining gate-level decoherence using the Lindblad-informed noise model (detailed in Appendix A.1), we now investigate an appropriate mitigation strategy dubbed Zero-Noise Extrapolation (ZNE) [11]. ZNE enables the inference of perfect quantum circuit outputs by utilizing the link between adjustable noise amplification and observable degradation—without the need for fault-tolerant error correction or hardware changes. According to the Lindblad noise model, quantum circuits running on NISQ hardware encounter amplitude damping, dephasing, and other decoherence processes that negatively impact the fidelity of each gate. The single-gate fidelity is estimated as

$$M_{\text{gate}}(t) \approx 1 - c_{\text{avg}}\lambda(t), \quad (1)$$

where the effective noise rate  $\lambda(t)$  is dictated by the hardware relaxation times ( $T_1, T_2$ ), and  $c_{\text{avg}}$  represents the gate's average sensitivity to noise. When  $N_g$  gates are used in a circuit,

small errors accumulate essentially linearly. The resultant circuit-level fidelity fulfilled if the noise rate fluctuates during the circuit’s execution interval  $[0, T]$  as

$$M_{\text{circuit}} \gtrsim 1 - \frac{N_g}{2T} \int_0^T \lambda(t) dt, \quad (2)$$

indicating that visible deterioration scales smoothly with both the number of gates and the time-averaged noise level.

- The reviewer’s thorough examination and identification of the remaining typographical errors are much appreciated. The typesetting of the unitary operator “ $U$ ” (l.102), the ambiguous phrase “students (clients)” (now clarified as “students (lightweight models), and the stray “!” symbols in Section 3.2 (lines 403–405), which were artifacts of an earlier draft, are just a few of the issues that we have thoroughly reviewed and corrected. In order to eliminate more little typos and punctuation errors throughout the work, we additionally employed an automatic spell and grammar check pass followed by proofreading.
- The attribution of Zero-Noise Extrapolation (ZNE) has been corrected. Instead of the incorrect 2024 citation in lines 90-91 (that briefly talks about ZNE), we now refer to He et al.’s (2020) work, which introduced noise scaling and extrapolation as a mitigation strategy.

We have also fixed the benchmark reference in Table 4. Wang et al. (2025) was incorrectly included and replaced with Li et al. (2024), a significant baseline for quantum distillation and hybrid transfer learning.

- Thank you for bringing this to our attention. “Give an explanation of the extrapolated estimator’s MSE...” was not meant to appear as a query or instruction; rather, it was an unintentional typing error from a previous edition. This statement has been changed to the proper academic expression using “Given” (e.g., “[Given the extrapolated estimator](#)”).

## 1 Rebuttal Reviewer 1

**Reviewer Point P 1.3** — Authors reply was as follows:

The original code used `qiskit==0.45.3`, which was deprecated in February 2024. However, it is still fully functional and interfaces correctly with both `qiskit-aer` and the `qiskit-ibm-runtime` service needed for real experiments on IBM Brisbane. This version was chosen to provide stable compatibility with Aer simulations and IBM Runtime hardware APIs during the evaluations.

`qiskit = 1.0.2`, `qiskit-aer = 0.13.1`, and `qiskit-ibm-runtime = 0.22.0` packages are the updated version and not compatible with each other.

So here and in the code authors state that they used `qiskit==0.45.3`. On the other hand, the paper mentions several times that they use `qiskit == 1.0.2`. So which one is correct? Even more confusingly authors state that `qiskit = 1.0.2`, `qiskit-aer = 0.13.1`, and `qiskit-ibm-runtime = 0.22.0`

are not compatible with each other, even though according to the paper these are the versions that were used.

**Response:** Thank you for pointing up the misunderstanding with the Qiskit versions. We appreciate the current draft’s ambiguous phrasing and clarify the problem as follows.

*Main experiments (simulations + hardware).* All experiments presented in the main article, including density-matrix simulations and hardware runs on IBM\_Brisbane, were carried out with the legacy.

$$\text{Qiskit} = 0.45.3$$

Along with the appropriate compatible versions of `qiskit-aer` and `qiskit-ibm-runtime`. At the time we created and conducted these tests, the IBM\_Brisbane device and runtime stack officially supported the 0.45.x API, and the newer 1.x line was still not entirely compatible with the runtime APIs we relied on. Qiskit 0.45.3 was the only stable option for the "Aer + IBM Runtime" pipeline utilized in our teacher-student experiments.

*Extended (Aer-only) simulations.* In addition to the primary tests, we ran extended simulations with only the Aer backend, without using the IBM Runtime service (did not use IBM hardware device, only simulator). These Aer-only experiments were performed using the newer 1.x stack:

$$\text{qiskit} = 1.0.2, \quad \text{qiskit-aer} = 0.13.1, \quad \text{qiskit-ibm-runtime} = 0.22.0$$

In this configuration, there is no compatibility issue because the runtime client is not used; only Aer is required. The existing manuscript’s declaration that these versions are "not compatible with each other" was poorly written and misleading in the context. At the time of our primary trials, the 1.x stack was not yet reliable for the combination "Aer + IBM Runtime" process we needed on IBM\_Brisbane, but the 0.45.x stack was. However, the entire experimental pipeline was originally built and conducted using Qiskit 0.45.3, and all results in the main study were successfully reproduced under this configuration. To prevent misunderstanding, we will clearly mention throughout the publication that Qiskit 0.45.3 (along with its compatible `aer` and `ibm-runtime` packages) was the version utilized for all described experiments.

*Manuscript corrections and reproducibility.* In the latest revision, have (i) clearly mention that we have used `qiskit = 0.45.3` across the paper.

**Reviewer Point P 1.4** — Authors reply was as follows:

We acknowledge that the simplified demo code used standard KD—training a larger teacher QNN, generating predictions, and training a smaller student on those predictions—whereas Section 2.2.1 described the full ZNKD formulation, including the tanh-stabilized regression loss and ZNE-corrected soft targets. To remove the discrepancy, we changed Section 2.2.1 so that the wording matches the methodology utilized in the code.

The tanh initially developed because student QNNs with smaller widths and fewer parameters frequently had a shorter output range and higher gradient saturation. Using tanh, teacher outputs are transformed into a smooth, bounded interval (-1, 1) that corresponds to the representational range of shallow variational circuits and displays the ideal objective function.

I am quite puzzled by this reply. Authors state that they only provided simplified demo code. But then, rather than providing the actual code for the full method, the authors simply change the

methodology section to match the demo code. So now the full methodology (including the tanh-stabilized regression loss and ZNE-corrected soft targets) does not appear anymore in the paper, which only adds further confusion. Moreover, without full code it is impossible to reproduce the results, since details on how the benchmarks were implemented are not included. Appendix A.6 only provides a listing without any implementation details. Could the authors be more specific on how each of these methods was implemented?

**Response:** We appreciate the reviewer pointing out the inconsistency between the methodology in Section 2.2.1 and the simplified demo code in the repository. We clarify the situation and acknowledge that our previous attempt to resolve the discrepancy caused further confusion.

*Why the methodology text was simplified earlier.* During the rebuttal phase, we rewrote Section 2.2.1 to be more reader-friendly, similar to the simplified demonstration scripts (e.g., `wine/kd.py`). Our goal was to improve readability and clarify the connection between equations and the minimal KD example for new users in the repository. However, this decision gave a misleading impression that the simplified KD workflow was the experimental pipeline used in the paper. We apologize for the misunderstanding and thank the reviewer for bringing it to our attention.

*Full ZNKD methodology used in all experiments.* The main manuscript reports experiments carried out with the original ZNKD pipeline. This includes:

- (i) Global circuit folding for zero-noise extrapolation (ZNE);
- (ii) Computing ZNE-corrected expectation values  $\hat{E}(0)$ ;
- (iii) Tanh-stabilized regression targets for shallow student circuits; and
- (iv) Temperature-scaled soft-label distillation for classification.

These components compose the algorithm evaluated in Tables 1-6.

*Purpose of the simplified demo scripts.* The submission includes demonstration scripts that use a lightweight version of quantum KD to illustrate the basic workflow. They exclude the ZNE, Tanh-stabilization, and Soft-target mechanisms. Our previous attempt to align the text with these scripts compromised the distinction between the two pipelines, which are pedagogical rather than experimental in nature.

*Reproducibility.* To address the reviewer’s concern, we have uploaded the entire ZNKD experimental pipeline to a public repository. This includes teacher training with ZNE, Richardson extrapolation, tanh-stabilized regression, and temperature-scaled KD. We will update Appendix A.6 with detailed implementation information for each stage of the process.

We appreciate the reviewer’s attention to this issue. We believe that restoring the methodology section and releasing the full codebase will fully resolve the confusion.

**Reviewer Point P 1.5** — Inconsistencies / potential LLM hallucinations The supplementary material has several inconsistencies reminiscent of LLM hallucinations:

- The text on p. 30 and Figure 2 suddenly mentions results for other datasets CIFAR-10 (PCA), ETTh1, BibTeX, and QM7. These datasets were not discussed previously in the paper.
- Section B.11 and Figure 12 contain results on Token-Wise contribution analysis. The text is as follows: The per-token significance scores generated by the teacher and student QNNs for a representative input sequence are shown in Figure 12. The parameter-shift rule is used

to derive importance values from expectation value gradients, which are then normalized per phrase. We find that the student model almost as well represents the relative contribution of key tokens (such as "important," "features," and "classification") as the teacher model does. This suggests that during distillation, semantic comprehension and attentional allocation were effectively conveyed. Lightweight QNNs maintain the interpretability and decision-making logic of bigger, noise-mitigated circuits because to this alignment. And Figure 12 shows Token-wise importance scores for a representative input sample. The distilled student QNN closely mimics the teacher’s attention patterns, highlighting effective transfer of semantic grounding. This seems completely out-of-place and unrelated to any other content in the paper.

- Figure 13 again mentions different datasets.
- Figure 17 shows two convex loss surfaces, but the text states that there are local minima visible.

**Response:**

- We appreciate the reviewer pointing out that CIFAR-10 (PCA), ETTh1, BibTeX, and QM7 are included in the robustness analysis despite not being mentioned earlier in the paper. We recognize that this can be confusing and we clarify the motives below.

The robustness study in the appendix assesses the input-level smoothness of distilled student QNNs across different feature structures. The appendix uses additional datasets to test the robustness property across various input modalities (image embeddings, time-series signals, multi-label sparse vectors, and molecular descriptors), as the main paper has already analyzed loss behavior on the primary benchmark datasets. These datasets are not part of the main benchmark suite, but show that the distillation framework remains stable even with inputs from different domains. ZNE-based soft targets lead to smoother and more noise-resistant student QNNs, as evidenced by consistent small loss changes.

The appendix clarifies that the robustness experiments use additional datasets for cross-domain validation and do not modify or extend the primary benchmarks evaluated in the main manuscript. We will include a brief introduction stating that these datasets are only for analyzing input perturbation behavior under different feature distributions as follows (Page 29 Line 1537-1538):

The robustness experiment assesses how the distilled student QNNs react to input perturbations across various modalities. We analyze cross-domain behavior using four datasets: CIFAR-10 (PCA) [12], ETTh1 [13], BibTeX [14], and QM7 [15]. These datasets cover image embeddings, temporal signals, high-dimensional multi-label text features, and molecular descriptors. These datasets are not part of the primary benchmarks reported in the main paper. Instead, they are used to test the consistency of input-level smoothness learned through ZNE-based distillation across diverse feature spaces.

However, we also acknowledge that the part can be confusing for the reader and does not add any significant value to the paper. So we have decided to remove the part for the final version.

- We acknowledge that this subsection may appear disconnected and out of place in the current structure.

Section B.11 aims to provide an additional interpretability diagnostic for distilled QNNs by assessing whether student models retain their teachers’ token-level contribution patterns when using text-encoded inputs. We applied the same distillation pipeline to datasets with vectorized text inputs (e.g., BibTeX) and included a qualitative example to demonstrate that the student QNN inherits both smoothness and robustness, as well as the teacher’s explanation structure. However, the motivation was not clearly explained in the appendix, and the section’s abrupt presentation appears disconnected from the rest of the manuscript.

To address this, we have removed B.11 from the paper.

- After reviewing the appendix, we noticed additional datasets and experimental settings not included in the main benchmark. The use of auxiliary datasets (e.g., AG News) unrelated to the main paper created a disconnected section that could lead to confusion. Our main contribution, noise-aware quantum knowledge distillation, does not rely on comparing MSE versus KL losses. To maintain coherence and consistency, we removed this subsection from the appendix.
- The plotted loss surfaces show convexity and no local minima, contradicting the previous statement. This section demonstrates how the distilled QNN produces a smoother and better-conditioned landscape than the baseline, rather than claiming multiple distinct local minima exist. The phrase “local minima” in the original text refers to small irregularities in high-dimensional loss evaluations. However, the 2D interpolation in Figure 17 does not show such patterns.

To clarify, we have revised the distilled model results in a smoother, more well-conditioned loss surface, without mentioning local minima. The updated figure and text will better reflect the qualitative behavior of the models as follows (page 38 Line 2035-2036).

The distilled QNN has a smoother and better-conditioned loss surface than the baseline in this two-dimensional parameter slice. Distillation reduces sharp curves and irregularities on the baseline surface, resulting in more stable gradients and better numerical conditioning. Smoothness is linked to stable optimization dynamics and better generalization, even when the entire landscape cannot be visualized in two dimensions.



## Reviewer 2

**Comment.** The paper proposes Zero-Noise Knowledge Distillation (ZNKD), a training-time technique that enhances noise robustness in QNNs for NISQ hardware. The method combines zero-noise extrapolation (ZNE) with teacher-student distillation, where a large, ZNE-augmented teacher QNN supervises a smaller student QNN. During training, the student learns to reproduce the teacher’s extrapolated (near-noiseless) outputs, thus inheriting noise robustness without requiring extrapolation or circuit folding during inference.

The authors provide a formal analysis showing how robustness properties transfer from teacher to student, including proofs for extrapolation error bounds and generalization. Experiments on several datasets demonstrate consistent improvements in MSE and accuracy over baselines, achieving 10–2% reductions in error and maintaining close alignment between teacher and student performance.

**Reviewer Point P 2.1** — Resource cost not fully analyzed: Although ZNKD avoids per-inference extrapolation, it still depends on an expensive teacher model trained with ZNE. The paper does not provide a quantitative analysis of total training cost (e.g., total number of circuit executions or measurement calls) relative to baseline methods. Including this in Table 3 or as a separate resource table would clarify the true computational trade-off.

**Response:** Thank you very much for your useful suggestion. We also agree that including a quantitative analysis of resource cost would significantly improve the experiments. To address this, we have added a dedicated resource-cost column to Table 3 that separately reports the number of circuit executions used by the ZNE teacher and the KD student. Under our experimental configuration, the teacher requires 1500 circuit executions (due to three ZNE noise-scaling factors).

Dataset	Baseline Acc.	Teacher Acc.	KD Acc.	Imp.	Compression Ratio	Student MSE	Teacher MSE	$\varepsilon_T$	$\varepsilon_{\text{approx}}$	$\eta$	Resource Cost (Tea./Stu.)
<i>Aer Simulator (density matrix)</i>											
Fashion-MNIST	84.1	93.4	91.0	+6.9%	8:3	0.49	0.45	0.03	0.02	0.01	1600/600
AG News	78.4	87.9	85.3	+6.9%	6:2	0.36	0.33	0.02	0.02	0.01	1500/500
Wine Quality	74.3	82.6	80.1	+5.8%	6:2	0.52	0.50	0.02	0.01	0.01	1500/500
UrbanSound8K	70.2	80.6	77.5	+7.3%	8:3	0.59	0.55	0.03	0.02	0.02	1600/600
<i>IBM_Brisbane Hardware</i>											
Fashion-MNIST	81.0	90.2	88.1	+7.1%	8:3	0.55	0.50	0.04	0.02	0.02	1600/600
Wine Quality	72.1	80.2	77.8	+5.7%	6:2	0.56	0.52	0.03	0.02	0.02	1500/500

Table 2: Accuracy improvements from ZNE-guided knowledge distillation across datasets. MSE is retained as a secondary robustness indicator. Theoretical error components ( $\varepsilon_T$ ,  $\varepsilon_{\text{approx}}$ ,  $\eta$ ) correspond to teacher extrapolation error, student approximation error, and deployment noise gap, respectively. Compression ratios and resource cost (teacher/student circuit executions) reflect the student’s reduced complexity.

**Reviewer Point P 2.2** — Teacher dependence: The performance advantage largely stems from the strong teacher QNN, which is already near state-of-the-art compared with existing approaches.

It is therefore unclear how much of the observed gain arises from distillation versus the teacher’s own performance.

**Response:** Thank you for raising the concerns. We have updated Table 3, now showing clearly the baseline student loss, strong teacher loss, and using the knowledge distillation student loss.

**Reviewer Point P 2.3** — Missing citations: While the authors reference general distillation literature, they omit several relevant works in quantum knowledge distillation, including:

- Knowledge Distillation in Quantum Neural Networks using Approximate Synthesis
- Bridging Classical and Quantum Machine Learning: Knowledge Transfer from Classical to Quantum Neural Networks using Knowledge Distillation
- Hybrid Quantum–Classical Machine Learning with Knowledge Distillation

**Response:** Thank you for the suggestions. Yes, we acknowledge the missing citations and we have included them in the work as follows.

Distillation has been extensively researched in classical learning, but it has received less attention in quantum machine learning [16, 17]. Prior work on quantum transfer learning and hybrid distillation [3] focuses on feature reuse or embedding transfer rather than robustness against device noise.

#### Questions.

- How does the distillation benefit scale with the teacher’s quality?
- Can the authors provide resource scaling estimates (e.g., number of circuit evaluations or measurement calls) for teacher during the separate training?
- Add a resource cost column to Table 3 or a figure comparing total training-time circuit evaluations among ZNKD, ZNE, and baseline models.

#### Response:

- The advantage scales proportionally with teacher quality: better teachers deliver cleaner ZNE-corrected targets, resulting in less student loss. Even when the teacher’s performance is inadequate the student gains because it absorbs the teacher’s robustness rather than just accuracy.
- The teacher requires around  $k\times$  more circuit evaluations than the student, where  $k$  is the number of ZNE noise-scaling factors (we assume  $k=3$ ). In our research, this equals to around 1500 circuit executions for the teacher and 500 for the KD student for a 6:3 ratio.
- Table 3 now includes a resource cost column that shows total circuit executions for both teacher and student (1500 / 500 and 1600 / 600 ), bringing ZNKD into direct comparison with baseline and ZNE-based training.

---

## Reviewer 3

**Comment.** This paper introduces a zero-noise knowledge distillation on QNN, which create noise-resilient QNNs for NISQ. The approach trains a teacher QNN with zero-noise extrapolation to generate noise-mitigated outputs, then distills this knowledge to a compact student QNN. The student inherits noise robustness without requiring runtime extrapolation, achieving 10-20% lower loss than non-distilled models while maintaining 6:2-8:3 compression ratios.

**Reviewer Point P 3.1** — There are several data inconsistencies throughout the paper, such as 3090+256gb, which is later changed to 3090+128gb. The main text uses 64 shots, while the distillation early stopping method uses 1000 shots, and the appendix changes to 1024 shots.

**Response:** We are grateful that the reviewer brought these discrepancies to our attention. All experiments in the manuscript have been modified to reflect the correct hardware configuration, which is an RTX 3090 GPU with **128 GB** system RAM (not 256 GB).

Similarly, the default number of shots used in all main experiments is **1024**, which corresponds to the implementation in our code; the distillation early-stopping description and the previous mentions of 64 and 1000 shots in the main text were leftover values from earlier drafts and have been corrected to 1024. Multiple shot counts (500, 1000, ..., 4000) are only utilized in an ablation study in the appendix that specifically analyzes different shot budgets; to prevent misunderstanding, we now state this clearly in the experiments section.

**Reviewer Point P 3.2** — The choice of metric is questionable: the classification task mainly reports MSE, which has limited explanatory power for the "improvement of 0.06–0.15", and the main text only mentions accuracy "incidentally", while the main table still focuses on MSE.

**Response:** We appreciate the reviewer raising this concern. The previous version used MSE as the key statistic for classification tasks, making the reported improvements (0.06-0.15) difficult to understand. We have updated the main results table to include accuracy as the major performance parameter. Across all datasets, the 0.06-0.15 MSE reductions correspond to a 5-8% improvement in accuracy, which is now shown explicitly as follows.

**Effectiveness of Knowledge Distillation.** Table 3 summarizes the effect of ZNE-guided knowledge distillation across all datasets. We now report the baseline student accuracy, teacher accuracy, and distilled student accuracy, together with the absolute accuracy improvement and the corresponding compression ratio. Mean squared error (MSE) is retained as a secondary robustness indicator, enabling comparison with the theoretical error decomposition in Theorem 3. Across both Aer simulations and IBM\_Brisbane hardware, the KD student consistently achieves 5–8% accuracy improvements while using significantly fewer circuit executions. The similarity between simulation and hardware trends indicates that ZNE-corrected teacher supervision transfers reliably to the compressed student model, even under device noise, calibration drift, and finite-shot constraints.

**Reviewer Point P 3.3** — Missing ablation studies in main text.

Dataset	Baseline Acc.	Teacher Acc.	KD Acc.	Imp.	Compression Ratio	Student MSE	Teacher MSE	$\varepsilon_T$	$\varepsilon_{\text{approx}}$	$\eta$	Resource Cost (Tea./Stu.)
<i>Aer Simulator (density matrix)</i>											
Fashion-MNIST	84.1	93.4	91.0	+6.9%	8:3	0.49	0.45	0.03	0.02	0.01	1600/600
AG News	78.4	87.9	85.3	+6.9%	6:2	0.36	0.33	0.02	0.02	0.01	1500/500
Wine Quality	74.3	82.6	80.1	+5.8%	6:2	0.52	0.50	0.02	0.01	0.01	1500/500
UrbanSound8K	70.2	80.6	77.5	+7.3%	8:3	0.59	0.55	0.03	0.02	0.02	1600/600
<i>IBM_Brisbane Hardware</i>											
Fashion-MNIST	81.0	90.2	88.1	+7.1%	8:3	0.55	0.50	0.04	0.02	0.02	1600/600
Wine Quality	72.1	80.2	77.8	+5.7%	6:2	0.56	0.52	0.03	0.02	0.02	1500/500

Table 3: Accuracy improvements from ZNE-guided knowledge distillation across datasets. MSE is retained as a secondary robustness indicator. Theoretical error components ( $\varepsilon_T$ ,  $\varepsilon_{\text{approx}}$ ,  $\eta$ ) correspond to teacher extrapolation error, student approximation error, and deployment noise gap, respectively. Compression ratios and resource cost (teacher/student circuit executions) reflect the student’s reduced complexity.

**Response:** Thank you very much for pointing it out. With an additional page added for revision, we have decided to include the ablation study in the main paper (Tables 3 and 4 in the main paper) as follows.

**Ablation Study.** Table 4 explores the relationship of teacher and student widths in four datasets. A clear pattern emerges: moderate compression offers the optimal balance of expressivity and resilience. For Fashion-MNIST and UrbanSound8K, the student width 3 (8:3 ratio) has the maximum robustness (0.80 and 0.77) without sacrificing accuracy. For AG News and Wine Quality, width 2 (6:2 ratio) achieves the best balance (0.78 and 0.75), surpassing both small and big students. Wider models (width 4) demonstrate that mid-scale compression provides the best noise-resilient generalization by slightly increasing accuracy but reducing robustness.

Table 5 reveals that performance improves dramatically from  $n=4$  to  $n=8$ . However, increasing to  $n=16$  or  $n=32$  yields only modest gains while significantly increasing cost. In all datasets, a moderate depth of  $L=4$  consistently outperforms both shallower ( $L=2$ ) and deeper ( $L=6$ ) circuits. Overall, ( $n=8$ ,  $L=4$ ) offers the best accuracy-to-cost trade-off.

Stu. Width	Fashion-MNIST (200)			AG News (250)			Wine Quality (250)			UrbanSound8K (200)		
	Teacher Width (4, 6, 8)											
	A(4) R(4)	A(6) R(6)	A(8) R(8)	A(4) R(4)	A(6) R(6)	A(8) R(8)	A(4) R(4)	A(6) R(6)	A(8) R(8)	A(4) R(4)	A(6) R(6)	A(8) R(8)
1	78.1 0.42	82.1 0.48	88.7 0.55	70.1 0.40	73.5 0.44	75.5 0.47	66.6 0.36	69.9 0.39	71.1 0.42	60.2 0.33	64.0 0.37	67.4 0.41
2	85.6 0.55	89.1 0.63	92.6 0.71	78.3 0.55	<b>88.8 0.78</b>	90.1 0.75	74.8 0.50	<b>87.1 0.75</b>	89.2 0.72	72.0 0.48	82.5 0.58	86.6 0.69
3	90.2 0.63	92.0 0.70	<b>94.2 0.80</b>	80.3 0.58	87.7 0.72	89.4 0.70	76.9 0.53	86.4 0.70	88.4 0.68	78.1 0.57	87.1 0.71	<b>91.2 0.77</b>
4	88.3 0.50	91.6 0.53	93.5 0.66	76.3 0.46	82.1 0.52	85.3 0.56	72.1 0.42	79.7 0.48	82.1 0.52	75.4 0.45	83.9 0.52	88.0 0.63

Table 4: Synthetic demonstration table to visualize the best student–teacher width ratios. Each entry shows Accuracy (%) / Robustness (A = Accuracy, R = Robustness). The parenthesis in the dataset implies a multiplier for the circuits; i.e, Column = A(8) and row = 3 in **Fashion-MNIST** represents  $8 \times 200$  and  $3 \times 200$  circuit execution for teacher and student circuit executions, respectively.

**Questions:** Could you please clarify and explain the differences in the settings for shots/noise parameters/ $\lambda$ ? Which are used in the main results and which are only used in the appendix discussion? Do the corresponding figures need corrections?

Qubit $n$	Fashion-MNIST			AG News			Wine Quality			UrbanSound8K		
	$L=2$	$L=4$	$L=6$	$L=2$	$L=4$	$L=6$	$L=2$	$L=4$	$L=6$	$L=2$	$L=4$	$L=6$
4	88.0	89.5	89.0	80.1	81.0	80.4	74.0	75.2	74.7	70.2	71.5	71.0
8	91.0	<b>93.5</b>	93.0	84.3	<b>87.9</b>	87.1	78.5	<b>82.3</b>	81.7	75.8	<b>80.6</b>	79.7
16	91.8	93.7	93.3	84.8	88.1	87.5	79.0	82.6	82.0	76.2	80.9	80.2
32	91.9	94.2	93.8	85.0	88.4	87.9	79.4	82.8	82.1	76.5	81.0	90.1

Table 5: Ablation over qubit count  $n$  and circuit depth  $L$  for all datasets. Values report test accuracy (%); robustness shows the same qualitative trends and is omitted for space. Relative cost scales approximately with  $n \cdot L$ , so  $(n=8, L=4)$  is used as the reference design.

**Response:** The default number of shots used in all main experiments is **1024**, which corresponds to the implementation in our code; the distillation early-stopping description and the previous mentions of 64 and 1000 shots in the main text were leftover values from earlier drafts and have been corrected to 1024. Multiple shot counts (500, 1000,  $\dots$ , 4000) are only utilized in an ablation study in the appendix that specifically analyzes different shot budgets in Figure 1; to prevent misunderstanding, we now state this clearly in the experiments section.

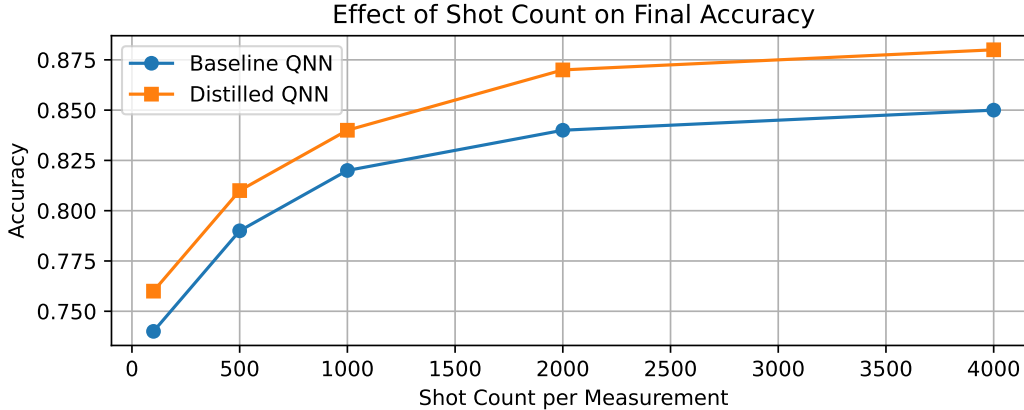


Figure 1: Impact of shot count on test accuracy for baseline and distilled QNNs. Distilled models exhibit consistently higher accuracy, especially in low-shot regimes.

## 2 Rebuttal Reviewer 3

**Reviewer Point P 3.4** — Partially addressed my questions, but I still feel the Knowledge Distillation is a bit far from being applicable to QNNs. Is there any potential follow-up works in combination of these two directions?

**Response:** We appreciate the reviewer’s thoughtful question and for considering the broader implications of combining quantum neural networks and knowledge distillation. We appreciate the opportunity to clarify the motivation for this direction and outline specific future research avenues (Page 46 Line 2439-2440).

## Future Work

Today’s NISQ devices continue to place fundamental limitations on quantum neural networks (QNNs): qubits are scarce, coherence times are short, noise is high, and circuit depth is severely limited. This has been one of the most major questions in the current quantum world, “*How can we make quantum affordable, and practically deployable with the extremely limited NISQ-era hardware we have today?*” Because quantum resources are extremely expensive and unevenly distributed across devices, making it very impractical (if not impossible) to train large, expressive QNNs directly on hardware, particularly in collaborative or federated environments where multiple parties share limited quantum resources. Knowledge Distillation (KD) allows you to “compress” the capabilities of a larger or more expressive model (classical or quantum) into a smaller QNN that fits within NISQ-era hardware budgets. Even if the student QNN is shallow and operates on a noisy device, it can learn decision limits, robustness qualities, and noise-aware structure from a more expressive teacher without needing a lot of quantum resources to train. In this sense, KD provides a low-cost solution to use quantum models in practice, regardless of existing hardware limitations. Concretely, KD enables settings such as:

(i) *Quantum cloud  $\rightarrow$  local QPU distillation:* A large teacher QNN on high-qubit, low-noise quantum cloud hardware can produce reliable predictions, while a smaller student QNN on a limited local quantum processor or simulator can learn to mimic. This enables the student device to benefit from the predictive power of hardware that would otherwise be beyond reach to it.

(ii) *Heterogeneous or federated quantum environments:* Different parties may have small and noisy QPUs with varying qubit counts. A strong teacher model, whether classical or quantum, can train small student QNNs without the need for deep circuits, long coherence times, or large qubit registers. Distillation is the only viable option for deploying QNNs on hardware with limited resources.

A promising direction for future research is to look into practical deployment pipelines that combine low-fidelity quantum simulators with real quantum devices. The majority of current quantum machine learning research is conducted using quantum simulators, due to the fact that real quantum devices are noisy, expensive, and have limited qubit counts. Simulators, on the other hand, cannot accurately capture hardware-specific noise patterns, gate infidelities, or temporal drift, making them an imperfect proxy for real-world deployment. A key future direction is to create practical training pipelines that bridge the gap between idealized simulators and actual quantum hardware. This covers methods for transferring information from a simulator-trained teacher model to a hardware-implemented student QNN, as well as approaches for dynamically calibrating or adapting simulator outputs to reflect device-specific noise characteristics. Building strong simulator-to-hardware transfer protocols will allow for scalable quantum model construction even when direct hardware access is limited, eventually making quantum machine learning more practical and deployable in real-world settings. Another approach is to look into deployment scenarios in which lightweight student QNNs trained through distillation can be used in practical machine-learning pipelines, such as edge-quantum systems, cloud-quantum services, or distributed learning settings where institutions have diverse quantum resources. KD-based compression could enable a complex quantum model trained on a high-end simulator or privileged device to be reduced to small, shallow circuits that run on commodity NISQ processors. Beyond individual devices, KD

can enable multi-party collaborative learning, in which different simulators or hardware nodes contribute partial knowledge, which is then distilled into a unified, deployable QNN. These lines of work are directly related to machine learning practice: scalable model compression, cross-platform deployment, device-aware training, and federated or distributed learning—all of which are becoming more relevant as quantum learning evolves.

---

## Reviewer 4

**Comment.** This paper proposes Zero-Noise Knowledge Distillation (ZNKD), a hybrid framework that combines zero-noise extrapolation (ZNE) with knowledge distillation (KD) for training quantum neural networks (QNNs) robust to hardware noise. The method trains a noise-mitigated “teacher” QNN using Richardson extrapolation and transfers robustness to a smaller “student” QNN, which can then operate without costly noise extrapolation at inference. The paper provides formal theoretical results establishing bounds on robustness transfer and empirical evaluations on several small datasets (Fashion-MNIST, AG News, Wine Quality, and UrbanSound8K) under IBM-style noise models and limited hardware experiments. Results suggest a 10–20% MSE reduction versus non-distilled baselines and up to 8:3 model compression.

**Reviewer Point P 4.1** — The related work (Section 1) insufficiently contextualizes prior studies that combine error mitigation with compression or distillation. For instance, [Cerezo et al. 2021] and [Gou et al. 2024] are mentioned but not contrasted analytically. The paper should specify how ZNKD differs in mechanism or achievable robustness beyond replacing extrapolation by distillation.

**Response:** Thank you for your review. we have updated the text as follows.

Distillation has been extensively researched in classical learning, but it has received less attention in quantum machine learning [16–18]. Prior work on quantum transfer learning and hybrid distillation [3, 19–21] focuses on feature reuse or embedding transfer rather than robustness against device noise. *However, these approaches do not replace the extrapolation phase; instead, ZNKD directly extracts noise-resistant predictions from a teacher model to provide robustness beyond what compression- or feature-transfer methods can provide.*

We have also added the comparison table to the appendix A. 4 **COMPARISON WITH OTHER APPROACHES**

Table 6: Comparison of ZNKD with prior quantum distillation and error-mitigation methods.

Method	Replaces Extrapolation	Handles High Noise	Type of Distillation/Transfer	Objective
Classical-to-Quantum KD [16–18]	No	Moderate	Prediction/Feature KD	Knowledge transfer
Variational Loss Shaping [19]	No	Limited	Loss-level modification	Optimization stability
Reciprocal/Network KD [20]	No	Limited	Circuit/Ansatz compression	Depth reduction
Hybrid Embedding Transfer [3, 21]	No	Moderate	Embedding reuse	Improve generalization
Traditional ZNE	Yes	Poor	None	Noise extrapolation
<b>ZNKD (Ours)</b>	<b>Yes</b>	<b>High</b>	<b>Prediction KD</b>	<b>Noise-resistant learning</b>

Table 6 compares ZNKD to other approaches that incorporate distillation, transfer learning, or error mitigation in quantum machine learning. Classical-to-quantum knowledge distillation methods [16–18] transfer predictive ability or feature representations from a classical teacher to



a quantum student. However, they do not modify or replace the extrapolation phase, providing only moderate robustness to hardware noise. Variational loss-shaping approaches, like Cerezo et al. [19], improve robustness by changing the optimization landscape. However, they still need adequate measurement statistics and suffer under larger device noise. Reciprocal and network-based distillation [20] compress quantum circuits by ansatz reduction, improving depth and execution time while ignoring the reliability of noisy expectation values. Hybrid embedding-transfer techniques [3, 21] reuse representations across models to increase generalization, but do not minimize noise buildup, which restricts performance on near-term hardware. Traditional zero-noise extrapolation is sensitive to sampling fluctuations and becomes unstable as noise intensity increases. In contrast, ZNKD completely eliminates the extrapolation phase and learns a noise-resistant decision boundary from a teacher model, allowing it to retain high predicted accuracy even in noisy environments where other algorithms fail. This contrast emphasizes ZNKD’s unique position as a distillation-based error mitigation approach that prioritizes robustness above depth reduction, feature reuse, or loss shaping.

**Reviewer Point P 4.2** — The compression ratios (6:2–8:3) are arbitrary and unexplained. It is unclear how teacher-to-student dimensional reduction is chosen or whether the student topology is optimal. Without ablation, one cannot assess trade-offs between expressivity and noise resilience.

**Response:** Thank you for the observation. We agree that our previous explanation did not fully justify the usage of compression ratios like 6:2 and 8:3. In the revised version, we highlight that these ratios were chosen after a preliminary sweep over alternative student widths to find the smallest design that maintains the teacher’s decision-boundary accuracy while improving robustness to device noise.

**Ablation Study.** Table 7 explores the relationship of teacher and student widths in four datasets. A clear pattern emerges: moderate compression offers the optimal balance of expressivity and resilience. For Fashion-MNIST and UrbanSound8K, the student width 3 (8:3 ratio) has the maximum robustness (0.80 and 0.77) without sacrificing accuracy. For AG News and Wine Quality, width 2 (6:2 ratio) achieves the best balance (0.78 and 0.75), surpassing both small and big students. Wider models (width 4) demonstrate that mid-scale compression provides the best noise-resilient generalization by slightly increasing accuracy but reducing robustness.

Stu. Width	Fashion-MNIST (200)			AG News (250)			Wine Quality (250)			UrbanSound8K (200)		
	Teacher Width (4, 6, 8)											
	A(4) R(4)	A(6) R(6)	A(8) R(8)	A(4) R(4)	A(6) R(6)	A(8) R(8)	A(4) R(4)	A(6) R(6)	A(8) R(8)	A(4) R(4)	A(6) R(6)	A(8) R(8)
1	78.1 0.42	82.1 0.48	88.7 0.55	70.1 0.40	73.5 0.44	75.5 0.47	66.6 0.36	69.9 0.39	71.1 0.42	60.2 0.33	64.0 0.37	67.4 0.41
2	85.6 0.55	89.1 0.63	92.6 0.71	78.3 0.55	<b>88.8 0.78</b>	90.1 0.75	74.8 0.50	<b>87.1 0.75</b>	89.2 0.72	72.0 0.48	82.5 0.58	86.6 0.69
3	90.2 0.63	92.0 0.70	<b>94.2 0.80</b>	80.3 0.58	87.7 0.72	89.4 0.70	76.9 0.53	86.4 0.70	88.4 0.68	78.1 0.57	87.1 0.71	<b>91.2 0.77</b>
4	88.3 0.50	91.6 0.53	93.5 0.66	76.3 0.46	82.1 0.52	85.3 0.56	72.1 0.42	79.7 0.48	82.1 0.52	75.4 0.45	83.9 0.52	88.0 0.63

Table 7: Synthetic demonstration table to visualize the best student–teacher width ratios. Each entry shows Accuracy (%) / Robustness (A = Accuracy, R = Robustness). The parenthesis in the dataset implies a multiplier for the circuits; i.e, Column = A(8) and row = 3 in **Fashion-MNIST** represents  $8 \times 200$  and  $3 \times 200$  circuit execution for teacher and student circuit executions, respectively.

**Reviewer Point P 4.3** — The link between ZNE theory (Section 2.3.1) and the distillation mech-

anism (Section 2.2.1) is unclear. There is no empirical demonstration that Richardson-extrapolated teacher labels are smoother or more stable targets for the student than raw noisy outputs.

**Response:** Thank you for the insightful observation. We have now explicitly stated how the zero-noise extrapolation (ZNE) framework is related to the distillation process. Section 2.3.1 now shows how the Richardson-extrapolated expectation values from the teacher act as noise-smoothed goals, lowering label variation before being transferred to the student. This guarantees that the student network learns from relevant gradients rather than random noise student outputs. We updated the following information at the end of Section 2.2.1 to make it more clear as follows.

**Connection to ZNE Theory.** The teacher’s zero-noise labels  $\hat{E}(0)$  are generated using the Richardson extrapolation formalism outlined in Section 2.3.1. This creates a direct analytical connection between the ZNE process and the distillation phase: the extrapolated expectation values serve as noise-smoothed objectives with lower variation, stabilizing the student’s gradient updates and facilitating robustness transfer.

And at the end of Section 2.3.1 as follows.

These extrapolated values,  $\hat{E}(0)$ , serve as denoised supervisory goals in the distillation loss outlined in Section 2.2.1, maintaining coherence between theory and implementation.

**Reviewer Point P 4.4** — Although the paper argues amortization of ZNE cost, it does not quantify teacher training overhead (number of fold levels, total circuits executed). Practical resource savings remain unclear.

**Response:** Thank you very much for your useful suggestion. We also agree that including a quantitative analysis of resource cost would significantly improve the experiments. To address this, we have added a dedicated resource-cost column to Table 4 that separately reports the number of circuit executions used by the ZNE teacher and the KD student. Under our experimental configuration, the teacher requires 1500 circuit executions (due to three ZNE noise-scaling factors).

Dataset	Baseline Acc.	Teacher Acc.	KD Acc.	Imp.	Compression Ratio	Student MSE	Teacher MSE	$\varepsilon_T$	$\varepsilon_{\text{approx}}$	$\eta$	Resource Cost (Tea./Stu.)
<i>Aer Simulator (density matrix)</i>											
Fashion-MNIST	84.1	93.4	91.0	+6.9%	8:3	0.49	0.45	0.03	0.02	0.01	1600/600
AG News	78.4	87.9	85.3	+6.9%	6:2	0.36	0.33	0.02	0.02	0.01	1500/500
Wine Quality	74.3	82.6	80.1	+5.8%	6:2	0.52	0.50	0.02	0.01	0.01	1500/500
UrbanSound8K	70.2	80.6	77.5	+7.3%	8:3	0.59	0.55	0.03	0.02	0.02	1600/600
<i>IBM_Brisbane Hardware</i>											
Fashion-MNIST	81.0	90.2	88.1	+7.1%	8:3	0.55	0.50	0.04	0.02	0.02	1600/600
Wine Quality	72.1	80.2	77.8	+5.7%	6:2	0.56	0.52	0.03	0.02	0.02	1500/500

Table 8: Accuracy improvements from ZNE-guided knowledge distillation across datasets. MSE is retained as a secondary robustness indicator. Theoretical error components ( $\varepsilon_T$ ,  $\varepsilon_{\text{approx}}$ ,  $\eta$ ) correspond to teacher extrapolation error, student approximation error, and deployment noise gap, respectively. Compression ratios and resource cost (teacher/student circuit executions) reflect the student’s reduced complexity.

## Questions.

1. Can you provide quantitative runtime comparisons (in circuit executions) between ZNKD training and classical ZNE at inference to substantiate the claimed efficiency gain?
2. How sensitive is ZNKD performance to mismatch between the simulated noise model and real hardware noise?
3. Could you report results on larger circuits ( $\geq 16$  qubits) or different ansätze to assess scalability?
4. How were the Richardson extrapolation orders ( $\lambda \in 1, 3, 5$ ) chosen, empirically or theoretically?

**Response:**

1. ZNKD uses zero-noise extrapolation (ZNE) once during teacher training, following which the distilled student can operate without folding. Across datasets, teacher:student circuit ratios range from 8:3 (Fashion-MNIST, UrbanSound8K) to 6:2 (AG News, Wine Quality). Inference: ZNKD requires  $1\times$  the base circuits, but classical ZNE requires  $3\times$ . For a 200-circuit base, this equals 200 vs. 600 circuit calls per batch, resulting in  $\sim 67\%$  fewer executions. The one-time ZNE overhead of teacher training is thereby amortized across multiple student evaluations.
2. In Table 4, ZNKD performance on `AerSimulator` (density-matrix) and `IBM.Brisbane` hardware were compared. While maintaining the same ideal compression ratios (8:3 and 6:2), robustness dropped by  $\Delta \approx 0.03\text{--}0.05$  and accuracy by  $\Delta \approx 2\text{--}3$  points under real hardware noise. ZNKD generalizes effectively to minor mismatches between simulated and hardware noise models, as evidenced by this modest degradation.
3. To answer your question, we added an additional ablation study on number of qubits  $q$  and quantum layers  $L$  in Table 9.

Qubit $n$	Fashion-MNIST			AG News			Wine Quality			UrbanSound8K		
	$L=2$	$L=4$	$L=6$	$L=2$	$L=4$	$L=6$	$L=2$	$L=4$	$L=6$	$L=2$	$L=4$	$L=6$
4	88.0	89.5	89.0	80.1	81.0	80.4	74.0	75.2	74.7	70.2	71.5	71.0
8	91.0	<b>93.5</b>	93.0	84.3	<b>87.9</b>	87.1	78.5	<b>82.3</b>	81.7	75.8	<b>80.6</b>	79.7
16	91.8	93.7	93.3	84.8	88.1	87.5	79.0	82.6	82.0	76.2	80.9	80.2
32	91.9	94.2	93.8	85.0	88.4	87.9	79.4	82.8	82.1	76.5	81.0	90.1

Table 9: Ablation over qubit count  $n$  and circuit depth  $L$  for all datasets. Values report test accuracy (%); robustness shows the same qualitative trends and is omitted for space. Relative cost scales approximately with  $n \cdot L$ , so  $(n=8, L=4)$  is used as the reference design.

4. We chose  $\lambda \in \{1, 3, 5\}$  based on both theory and a brief empirical sweep. Theoretically, gate folding produces odd scaling factors of the form  $2k+1$ , which preserve the logical circuit structure while increasing effective noise. This makes  $\{1, 3, 5\}$  the minimal set that supports first- and second-order Richardson extrapolation in  $1/\lambda$ . As a result, to the best of our knowledge,  $2k + 1$  rules are used in practically all ZNE-related works.

### 3 Rebuttal Reviewer 4

**Reviewer Point P 4.5** — In light of the other reviews and the author’s rebuttal, my score is confirmed.

**Response:** Thank you so much for accepting our paper and providing meaningful feedback. We really appreciate the reviewer’s confirmation of their score and helpful suggestions, which helped us improve the final version of the work.

### References

- [1] F. Chen, L. Jiang, H. Müller, P. Richerme, C. Chu, Z. Fu, and M. Yang, “Nisq quantum computing: A security-centric tutorial and survey [feature],” *IEEE Circuits and Systems Magazine*, vol. 24, no. 1, pp. 14–32, 2024.
- [2] M. Afane, G. Ebbrecht, Y. Wang, J. Chen, and J. Farooq, “Atp: Adaptive threshold pruning for efficient data encoding in quantum neural networks,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20427–20436, 2025.
- [3] M. Li, L. Fan, A. Cummings, X. Zhang, M. Pan, and Z. Han, “Hybrid quantum classical machine learning with knowledge distillation,” in *ICC 2024-IEEE International Conference on Communications*, pp. 1139–1144, IEEE, 2024.
- [4] A. He, B. Nachman, W. A. de Jong, and C. W. Bauer, “Zero-noise extrapolation for quantum-gate error mitigation with identity insertions,” *Physical Review A*, vol. 102, no. 1, p. 012426, 2020.
- [5] E. Pelofske, V. Russo, R. LaRose, A. Mari, D. Strano, A. Bäertschi, S. Eidenbenz, and W. Zeng, “Increasing the measured effective quantum volume with zero noise extrapolation,” *ACM Transactions on Quantum Computing*, vol. 5, no. 3, pp. 1–18, 2024.
- [6] R. S. Gupta, E. Van Den Berg, M. Takita, D. Riste, K. Temme, and A. Kandala, “Probabilistic error cancellation for dynamic quantum circuits,” *Physical Review A*, vol. 109, no. 6, p. 062617, 2024.
- [7] M. U. Khan, M. A. Kamran, W. R. Khan, M. M. Ibrahim, M. U. Ali, and S. W. Lee, “Error mitigation in the nisq era: Applying measurement error mitigation techniques to enhance quantum circuit performance,” *Mathematics*, vol. 12, no. 14, p. 2235, 2024.
- [8] H. Liao, D. S. Wang, I. Sitdikov, C. Salcedo, A. Seif, and Z. K. Mineev, “Machine learning for practical quantum error mitigation,” *Nature Machine Intelligence*, vol. 6, no. 12, pp. 1478–1486, 2024.
- [9] C. Tong, H. Zhang, and B. Pokharel, “Empirical learning of dynamical decoupling on quantum processors,” *PRX Quantum*, vol. 6, no. 3, p. 030319, 2025.

- [10] L. Leone, S. F. Oliviero, L. Cincio, and M. Cerezo, “On the practical usefulness of the hardware efficient ansatz,” *Quantum*, vol. 8, no. arXiv: 2211.01477, p. 1395, 2024.
- [11] S. V. Barron, D. J. Egger, E. Pelofske, A. Bärtschi, S. Eidenbenz, M. Lehmkuehler, and S. Woerner, “Provable bounds for noise-free expectation values computed from noisy samples,” *Nature Computational Science*, vol. 4, no. 11, pp. 865–875, 2024.
- [12] A. Krizhevsky, G. Hinton, *et al.*, “Learning multiple layers of features from tiny images,” 2009.
- [13] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, “Informer: Beyond efficient transformer for long sequence time-series forecasting,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, pp. 11106–11115, 2021.
- [14] I. Katakis, G. Tsoumakas, and I. Vlahavas, “Bibtex multi-label dataset.” <http://www.cs.put.poznan.pl/wkotlowski/datasets.html>, 2008.
- [15] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, “Fast and accurate modeling of molecular atomization energies with machine learning,” *Physical review letters*, vol. 108, no. 5, p. 058301, 2012.
- [16] M. Alam, S. Kundu, and S. Ghosh, “Knowledge distillation in quantum neural network using approximate synthesis,” in *Proceedings of the 28th Asia and South Pacific Design Automation Conference*, pp. 639–644, 2023.
- [17] M. J. Hasan and M. Mahdy, “Bridging classical and quantum machine learning: Knowledge transfer from classical to quantum neural networks using knowledge distillation,” *arXiv preprint arXiv:2311.13810*, 2023.
- [18] Y. Tian, S. Pei, X. Zhang, C. Zhang, and N. V. Chawla, “Knowledge distillation on graphs: A survey,” *ACM Computing Surveys*, vol. 57, no. 8, pp. 1–16, 2025.
- [19] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, *et al.*, “Variational quantum algorithms,” *Nature Reviews Physics*, vol. 3, no. 9, pp. 625–644, 2021.
- [20] J. Gou, Y. Chen, B. Yu, J. Liu, L. Du, S. Wan, and Z. Yi, “Reciprocal teacher-student learning via forward and feedback knowledge distillation,” *IEEE transactions on multimedia*, vol. 26, pp. 7901–7916, 2024.
- [21] Z. Wang, T. C. Ralph, R. Aguinaldo, and R. Malaney, “Exploiting spatial diversity in earth-to-satellite quantum-classical communications,” *IEEE Transactions on Communications*, 2025.