Rebuttal for ICML'25 Submission #3922

Table 1: **Response to Q4 and Q6 of Reviewer HcSt, Q3 of Reviewer Wxrf, with respect to the question about how the performance of deeper networks changes.** The RMSE result on correlation prediction of $|\psi_{\mathrm{HB}}\rangle$ with varied system size N and finetuning training size. M is fixed to 64. `MLP(CNN)-x layers` represents neural network `MLP` (`CNN`) that composed of `x` layers with residual connection. The best results are highlighted in **boldface** while the second-best results are distinguished in underlined. As networks go deeper, performance on predicting $\bar{C}$ of $|\psi_{\mathrm{HB}}\rangle$ improves then declines, yet still is inferior to classical ML models.

Methods	N = 48			N = 63			N = 100			N = 127		
	$n_{\mathrm{sft}}=20$	$n_{\mathrm{sft}}=60$	$n_{\mathrm{sft}}=100$	$n_{\mathrm{sft}}=20$	$n_{\mathrm{sft}}=60$	$n_{\mathrm{sft}}=100$	$n_{\mathrm{sft}}=20$	$n_{\mathrm{sft}}=60$	$n_{\mathrm{sft}}=100$	$n_{\mathrm{sft}}=20$	$n_{\mathrm{sft}}=60$	$n_{\mathrm{sft}}=100$
CS	0.21113			0.21257			0.21399			0.21447		
MLP-2 layers	0.08282	0.07752	0.06616	0.12055	0.08776	0.07086	0.10848	0.08158	0.07405	0.10091	0.10083	0.08245
MLP-3 layers	0.06214	0.04853	0.04494	0.07256	0.05506	0.04467	0.07740	0.06496	0.07098	0.08535	0.08280	0.08691
MLP-4 layers	0.05428	0.03825	0.03524	0.06463	0.04435	0.03833	0.07532	0.05952	0.06010	0.07971	0.09173	0.08608
MLP-5 layers	0.07228	0.04721	0.03764	0.07308	0.05957	0.05091	0.08046	0.07146	0.07174	0.08408	0.08650	0.08458
CNN-2 layers	0.07160	0.04723	0.03795	0.07176	0.04066	0.03042	0.06549	0.04566	0.03464	0.06468	0.03189	0.07404
CNN-3 layers	0.08089	0.03422	0.03435	0.09003	0.03401	0.03159	0.07603	0.03245	0.03295	0.08420	0.03179	0.03025
CNN-4 layers	0.06484	0.04899	0.03456	0.06621	0.03608	0.03100	0.06436	0.03425	0.02808	0.07441	0.03196	0.05221
CNN-10 layers	0.06388	0.08577	0.03856	0.13669	0.06697	0.09836	0.05456	0.03361	0.03555	0.05273	0.08775	0.03523
CNN-20 layers	0.15740	0.11951	0.07480	0.13665	0.10532	0.07100	0.11759	0.09031	0.07029	0.10187	0.08780	0.07183
CNN-50 layers	0.16392	0.12271	0.07735	0.16071	0.14676	0.09655	0.14741	0.11789	0.09367	0.13320	0.12921	0.10086
CNN-100 layers	0.20797	0.20659	0.20394	0.18382	0.17760	0.17323	0.14762	0.14402	0.13628	0.13455	0.13356	0.13150
LLM4QPE-T	0.05189	0.03368	0.03197	0.06111	0.03364	0.02863	0.05050	0.03227	0.02726	0.05079	0.03184	0.02634
RBFK	0.05452	0.04176	0.04101	0.04726	0.03829	0.03922	**0.04096**	0.03299	0.03282	**0.03850**	0.03115	0.03086
Lasso	**0.04221**	**0.02636**	0.02489	0.04856	**0.02791**	**0.02326**	0.04219	**0.02602**	0.02646	0.04137	0.03292	**0.02083**
Ridge	0.04247	0.02884	**0.02475**	**0.04216**	0.02816	0.02402	0.04191	0.02711	**0.02251**	0.04110	**0.02620**	0.02161

Table 2: **Response to Q4 and Q6 of Reviewer HcSt, Q3 of Reviewer Wxrf, with respect to the question about how the performance of deeper networks changes.** The RMSE result on correlation prediction of $|\psi_{\mathrm{TFIM}}\rangle$ with varied system size N and finetuning training size n_{sft}. M is fixed to 64. `MLP(CNN)-x layers` represents neural network `MLP` (`CNN`) that composed of `x` layers with residual connection. The best results are highlighted in **boldface** while the second-best results are distinguished in underlined. As networks go deeper, performance on predicting $\bar{C}$ of $|\psi_{\mathrm{TFIM}}\rangle$ improves then declines, yet still is inferior to classical ML models.

Methods	N = 48			N = 63			N = 100			N = 127		
	$n_{\mathrm{sft}}=20$	$n_{\mathrm{sft}}=60$	$n_{\mathrm{sft}}=100$	$n_{\mathrm{sft}}=20$	$n_{\mathrm{sft}}=60$	$n_{\mathrm{sft}}=100$	$n_{\mathrm{sft}}=20$	$n_{\mathrm{sft}}=60$	$n_{\mathrm{sft}}=100$	$n_{\mathrm{sft}}=20$	$n_{\mathrm{sft}}=60$	$n_{\mathrm{sft}}=100$
CS	0.20924			0.20990			0.21092			0.21180		
MLP-2 layers	0.07899	0.06371	0.05524	0.07986	0.05279	0.04283	0.08293	0.05303	0.04630	0.07908	0.05006	0.04333
MLP-3 layers	0.06080	0.05664	0.06074	0.06514	0.06928	0.06914	0.06301	0.06358	0.07317	0.06324	0.06510	0.07327
MLP-4 layers	0.05912	0.05794	0.05980	0.05899	0.05705	0.06163	0.05678	0.05628	0.06977	0.05535	0.06496	0.07197
MLP-5 layers	0.07422	0.06545	0.05739	0.07341	0.06921	0.069215	0.06648	0.06556	0.07044	0.06941	0.07222	0.06867
CNN-2 layers	0.12845	0.15039	0.08935	0.12227	0.16686	0.10315	0.10084	0.08879	0.05177	0.10495	0.08535	0.04647
CNN-3 layers	0.13545	0.17135	0.12004	0.12545	0.17026	0.11778	0.11433	0.11267	0.05027	0.13312	0.03562	0.05347
CNN-4 layers	0.13624	0.17178	0.12015	0.12608	0.17103	0.13809	0.12221	0.11046	0.06586	0.13757	0.10498	0.05556
CNN-10 layers	0.10861	0.14012	0.13969	0.10894	0.14113	0.13640	0.08386	0.10294	0.06330	0.07107	0.06095	0.04910
CNN-20 layers	0.06796	0.07030	0.09552	0.05565	0.03468	0.03917	0.17534	0.10762	0.04129	0.05152	0.03588	0.04086
CNN-50 layers	0.05984	0.03783	0.20409	0.29550	0.27408	0.23003	0.27766	0.03706	0.04305	0.28359	0.26455	0.22790
CNN-100 layers	0.31863	0.31729	0.31449	0.31156	0.31115	0.30988	0.30174	0.30136	0.30013	0.29768	0.29570	0.29139
LLM4QPE-T	0.05088	0.03493	0.03006	0.05252	0.03566	0.03082	0.05217	0.03476	0.03012	0.05259	0.03641	0.03084
Lasso	0.04624	0.03219	0.02812	0.04633	0.03930	0.02859	**0.04073**	0.03256	0.02899	0.04583	**0.03283**	0.02932
Ridge	**0.04473**	**0.03173**	**0.02807**	**0.04561**	**0.03226**	**0.02839**	0.04598	0.03277	**0.02883**	**0.04570**	0.03285	**0.02911**

Table 3: **Response to Q2 and Q5 of Reviewer HcSt, Q3 of Reviewer 3fBm, with respect to the question about how the performance of models if considering a larger amount of data changes under numerical simulation settings.** The RMSE results on correlation prediction of $|\psi_{\mathrm{HB}}\rangle$ with varied N. The training set and testing set both have 10^4 samples, with noise-free labels ($M \to \infty$). The best results are highlighted in **boldface**. As training data amounts expand (at most $\times 100$) and considering infinite measurement shots, the performance of `Ridge` on predicting $\bar{C}$ of $|\psi_{\mathrm{HB}}\rangle$ is superior to that of other advance DL models, varied system size from 8 to 31.

$M \to \infty$	$N = 8$	$N = 10$	$N = 12$	$N = 16$	$N = 25$	$N = 31$
Ridge	**0.00367**	**0.00444**	**0.00566**	**0.00636**	**0.00599**	**0.00579**
MLP-4 layers	0.03961	0.03677	0.03460	0.03129	0.02769	0.02625
CNN-4 layers	0.02056	0.03710	0.03432	0.03050	0.02582	0.02381
LLM4QPE-F	0.04666	0.04385	0.03969	0.03728	0.03083	0.02951

Table 4: **Response to Q2 and Q5 of Reviewer HcSt, Q3 of Reviewer 3fBm, with respect to the question about how the performance of models if considering a larger amount of data changes under numerical simulation settings.** The RMSE results on predicting correlation of $|\psi_{\mathrm{HB}}\rangle$ with varied training size n. System size $N = 8$. The number of testing sets is fixed to 2×10^4. Labels are noise-free ($M \to \infty$). The best results are highlighted in **boldface**. As training data amounts expand (at most $\times 1000$) and considering infinite measurement shots, the performance of `Ridge` on predicting $\bar{C}$ of 8-qubit $|\psi_{\mathrm{HB}}\rangle$ is superior to that of other advance DL models.

$M \to \infty$	# Params.	$n = 10^2$	$n = 10^3$	$n = 10^4$	$n = 10^5$
Ridge	< 0.01M	**0.00780**	**0.00528**	**0.00367**	**0.00660**
MLP-4 layers	0.09M	0.04219	0.04172	0.03961	0.03956
CNN-4 layers	1.14M	0.01987	0.02078	0.02056	0.02054
LLM4QPE-F	9.89M	0.03966	0.04304	0.04916	0.04659

Table 5: **Response to Q2 and Q5 of Reviewer HcSt, Q3 of Reviewer 3fBm, with respect to the question about how the performance of models if considering a larger amount of data changes under numerical simulation settings.** The RMSE results on predicting entanglement entropy of $|\psi_{\mathrm{HB}}\rangle$ with varied training size n. System size $N = 8$. The number of testing sets is fixed to 2×10^4. Labels are noise-free ($M \to \infty$). The best results are highlighted in **boldface**. As training data amounts expand (at most $\times 1000$) and considering infinite measurement shots, the performance of `Ridge` on predicting $\bar{S}_2$ of 8-qubit $|\psi_{\mathrm{HB}}\rangle$ is superior to that of other advance DL models.

$M \to \infty$	# Params.	$n = 10^2$	$n = 10^3$	$n = 10^4$	$n = 10^5$
Ridge	< 0.01M	**0.01563**	**0.00947**	**0.00753**	**0.00851**
ResMLP-2 layers	0.09M	0.10817	0.09142	0.05398	0.05302
ResCNN-4 layers	1.14M	0.04334	0.02410	0.03520	0.02073
LLM4QPE-F	9.89M	0.10648	0.11171	0.10895	0.10826

Table 6: **Response to Q1 of Reviewer 3fBm with respect to the question about how the performance changes with the role of embeddings.** The RMSE results of `LLM4QPE-F` on correlation prediction of N-qubit $|\psi_{\mathrm{HB}}\rangle$, with embedding M_{emb} random measurement outcomes. The training set and testing set both have 10^4 samples, with noise-free labels ($M \to \infty$). M_{emb} is the actual number of embedded measurement outcomes. As the number of embedded random outcomes increases, the performance of `LLM4QPE` decreases.

	$N = 8$	$N = 10$	$N = 12$	$N = 16$	$N = 25$	$N = 31$
$M_{\mathrm{emb}} = 1$	0.04666	0.04385	0.04126	0.03728	0.03083	0.03125
$M_{\mathrm{emb}} = 8$	0.04746	0.04926	0.03969	0.03984	0.03408	0.02951
$M_{\mathrm{emb}} = 64$	0.04795	0.04791	0.04785	0.04043	0.03637	0.03524
$M_{\mathrm{emb}} = 512$	0.04913	0.04521	0.04506	0.03905	0.03406	0.03268

Table 7: **Response to Q1 of Reviewer 3fBm with respect to the question about how the performance changes with the role of embeddings.** The RMSE results of `LLM4QPE-F` on correlation prediction of N-qubit $|\psi_{\mathrm{HB}}\rangle$, with embedding M_{emb} real measurement outcomes over the finetuning phase. testing size is set to 200. M is fixed to 512. $M_{\mathrm{emb}} \leq M$ is the actual number of embedded measurement outcomes. n_{sft} is the training size over the finetuning phase. As the actual number of real outcomes is embedded in the model, the performance of `LLm4QPE` remains the same, which reinforces that the LLM-like embedding approach makes the outcomes redundant features.

| | $N = 63$ | | | $N = 100$ | | | $N = 127$ | | |
	$n_{\mathrm{sft}} = 20$	$n_{\mathrm{sft}} = 60$	$n_{\mathrm{sft}} = 100$	$n_{\mathrm{sft}} = 20$	$n_{\mathrm{sft}} = 60$	$n_{\mathrm{sft}} = 100$	$n_{\mathrm{sft}} = 20$	$n_{\mathrm{sft}} = 60$	$n_{\mathrm{sft}} = 100$
$M_{\mathrm{emb}} = 1$	0.02555	0.02104	0.02019	0.02307	0.01872	0.01760	0.02239	0.01739	0.01635
$M_{\mathrm{emb}} = 8$	0.02556	0.02106	0.02019	0.02309	0.01873	0.01760	0.02242	0.01739	0.01635
$M_{\mathrm{emb}} = 64$	0.02556	0.02104	0.02019	0.02309	0.01872	0.01759	0.02239	0.01739	0.01636
$M_{\mathrm{emb}} = 512$	0.02560	0.02104	0.02019	0.02309	0.01872	0.01759	0.02240	0.01740	0.01635

Table 8: **Response to Q1 of Reviewer 3fBm and Q2 of Review Wxrf, with respect to the question about how the performance changes with the role of embeddings.** The RMSE results of predicting correlation of N-qubit $|\psi_{\mathrm{HB}}\rangle$, with MLP, `Lasso` and `Ridge` as learning models. Measurement outcomes are embedded as input features of MLP in two ways: `raw` tensor directly characterizing, or averaging (`Avg.`) over M measurement outcomes for each qubit ($M \times N \to 1 \times N$). $N \in \{63, 100, 127\}$. Training size $n \in \{20, 80, 100\}$. Measurement shots $M \in \{64, 128, 256, 512\}$. Simply averaging over outcomes for each qubit could significantly increase the performance of MLP, yet it is still inferior to `Lasso` and `Ridge`.

| | | $N = 63$ | | | $N = 100$ | | | $N = 127$ | | |
		$n = 20$	$n = 60$	$n = 100$	$n = 20$	$n = 60$	$n = 100$	$n = 20$	$n = 60$	$n = 100$
M=64	Raw	0.08964	0.05522	0.04872	0.08666	0.04949	0.04055	0.08878	0.05068	0.04076
	Avg.	0.05572	0.03522	0.02984	0.05525	0.03972	0.02801	0.05505	0.03951	0.03242
	Lasso	0.04856	0.02791	0.02326	0.04219	0.02602	0.02646	0.04137	0.03292	0.02083
	Ridge	0.04216	0.02816	0.02402	0.04191	0.02711	0.02251	0.04110	0.02620	0.02161
M=128	Raw	0.10921	0.05905	0.04835	0.10966	0.06137	0.04485	0.10408	0.06359	0.04554
	Avg.	0.04403	0.03034	0.02552	0.04699	0.03561	0.02603	0.04435	0.03421	0.03007
	Lasso	0.03168	0.02171	0.01905	0.03127	0.02045	0.01735	0.03041	0.01980	0.01647
	Ridge	0.03169	0.02178	0.01921	0.03069	0.02067	0.01786	0.03053	0.02087	0.01726
M=256	Raw	0.14085	0.08316	0.06045	0.12558	0.08648	0.05983	0.11720	0.08232	0.06089
	Avg.	0.03581	0.02673	0.02272	0.04022	0.02966	0.02168	0.03883	0.03188	0.02893
	Lasso	0.02556	0.01749	0.12125	0.02406	0.01747	0.01467	0.02283	0.01542	0.01324
	Ridge	0.02556	0.01751	0.01572	0.02408	0.01697	0.01494	0.02286	0.01576	0.01377
M=512	Raw	0.15943	0.11187	0.08246	0.13586	0.10826	0.08329	0.12608	0.10324	0.08330
	Avg.	0.03020	0.02475	0.02211	0.03713	0.02864	0.02272	0.03644	0.02962	0.02618
	Lasso	0.02037	0.01586	0.11038	0.01892	0.01403	0.01263	0.01702	0.01257	0.01117
	Ridge	0.02036	0.01583	0.01436	0.01891	0.01404	0.01271	0.01798	0.01285	0.01186

Table 9: **Response to Q4 of Reviewer 3fBm with respect to the question about the reverse phenomenon of two datasets for the same task in Fig.1 of our original manuscript.** The RMSE results of `Ridge` on predicting correlation of N-qubit $|\psi_{\mathrm{HB}}\rangle$ and $|\psi_{\mathrm{TFIM}}\rangle$. The input dimension d is both fixed to 20. Regularization of Ridge is set to $\lambda = 1$. The performance `Ridge` on predicting $\bar{C}$ of $|\psi_{\mathrm{HB}}$ and $|\psi_{\mathrm{TFIM}}$ exhibits comparable results, if fix the model input the same.

| Dataset | $N = 63$ | | | | | $N = 100$ | | | | | $N = 127$ | | | | |
	$n = 20$	$n = 40$	$n = 60$	$n = 80$	$n = 100$	$n = 20$	$n = 40$	$n = 60$	$n = 80$	$n = 100$	$n = 20$	$n = 40$	$n = 60$	$n = 80$	$n = 100$
HB	**0.09998**	0.10555	0.09941	0.09322	0.08782	**0.10015**	**0.10395**	0.09867	0.09278	**0.08692**	**0.09964**	0.10491	**0.09898**	**0.09241**	**0.08680**
TFIM	0.10185	**0.10333**	**0.09845**	**0.09189**	**0.08565**	0.10093	0.10436	**0.09847**	**0.09193**	0.08824	0.10148	**0.10372**	0.10106	0.09426	0.08716