IONQ

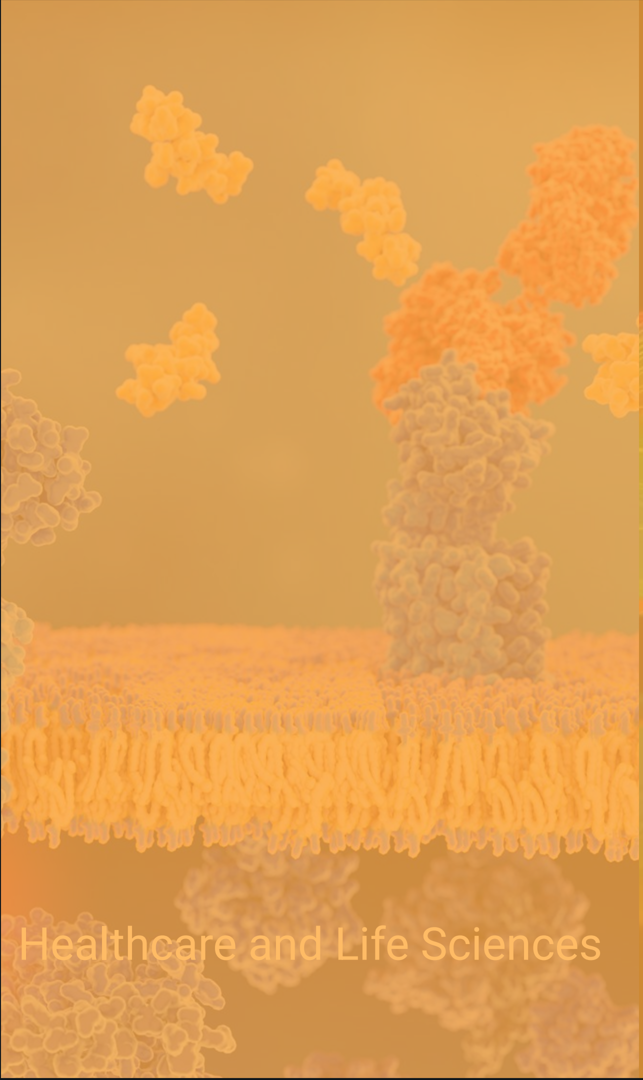
Hybrid Quantum-Classical Architecture for Large Language Model Fine-Tuning: Toward Hybrid CPU + GPU + QPU

ICS2025

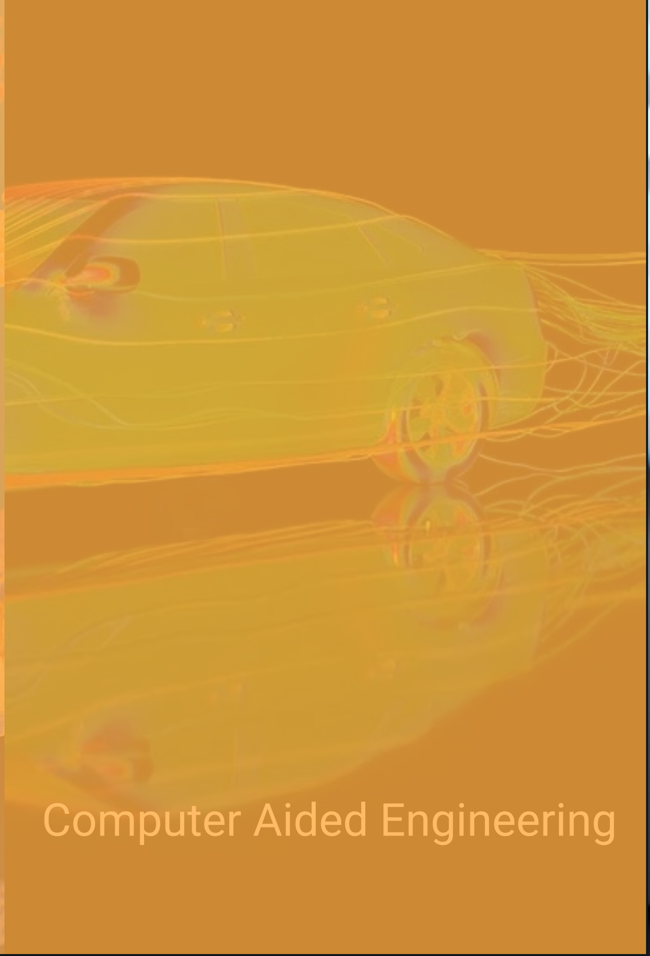
June 8, 2025

Sang Hyub Kim, Jonathan Mei, **Claudio Girotto**, Masako Yamada, Martin Roetteler
Technical Program Manager Applications, IonQ

Quantum Is Now



Healthcare and Life Sciences



Computer Aided Engineering

Quantum + AI

Overview

- Hybrid quantum-classical deep learning architecture for large language model fine-tuning.
- Performance for various settings of hyperparameters
- Prediction accuracy increasing with number of qubits, improving over a comparable classical baseline

Examples of Foundation AI Fine-Tuning

Original purpose of the Base (Large Language) Model (LLM):

Open-source general language generation.

Mistral → Customer Support Agent

GPT-J → Mental Health Companion

DeepSeek R1 → Planning

SetFit / BERT (this work) → Sentiment prediction

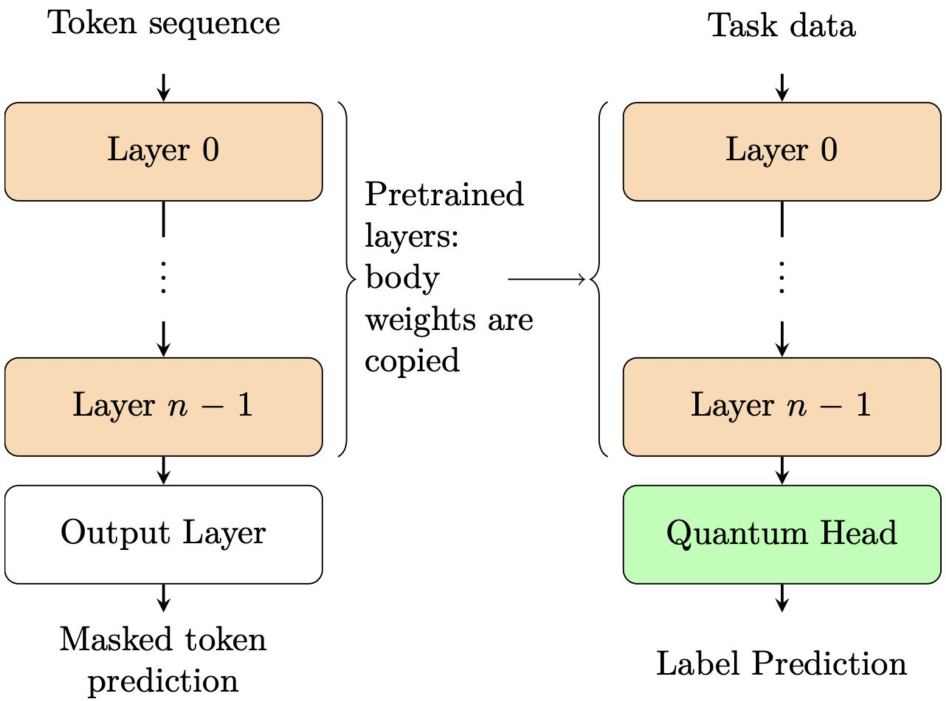
Different strategies in LLM improvement:

- **Fine-tuning**
- **Retrieval-Augmented Generation**
- **Knowledge Graphs**
- **Agentic AI**

Quantum + AI: Fine-tuning to teach LLMs new tasks

"The quick [masked] fox jumped…"

"This movie doesn't care about humor, wit,…"

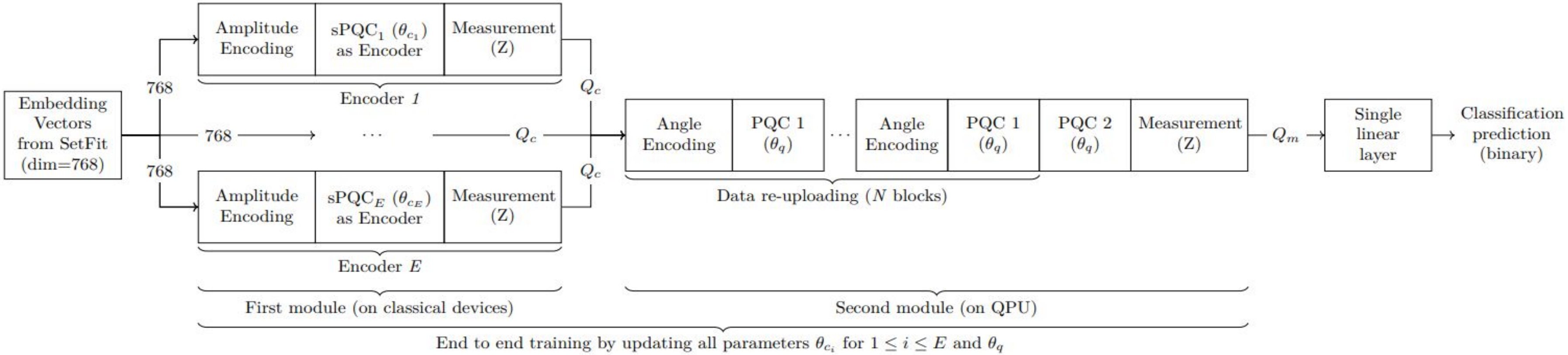


"The quick *red* fox jumped…"

negative sentiment

https://arxiv.org/abs/2504.08732

Quantum + AI fine-tuning: hybrid architecture



Our novel hybrid machine learning model combines classical computing with a Quantum Processing Unit (QPU) to reduce the input size and classify data - all trained together in one system.

- The SetFit Embeddings do the initial text embedding
- The Classical Module takes care of pre-processing and hybrid encoding
- The Quantum Module is the Quantum Encoder + Quantum Head
- The Linear Classifier is the optional final layer after quantum output
- The Binary Prediction layer makes the final prediction.

Quantum + AI fine-tuning: hyperparameters considered

Hyperparameter	Symbol	Example (exploring encoder)
Qubits	Q	10, 12, 14, 16, 18
Number of Encoders	E	4
Re-upload number	R	2
Number of Main blocks	M	1
Number of re-uploading blocks	N	1
Connectivity	C	16
Batch Size	B	8192
Number of Shots	S	Yes
Final linear layer	F	$\{1, 1.5, 2, 2.5, 3, 5\}/10^3$
Learning rate	L	1.0
Learning rate decay	γ	0.0
Weight decay	ρ	

Compute setup

We performed experiments using 3 NVIDIA GPU types:

- L4 (24GB)
- A100 (80GB)
- H100 (80GB)

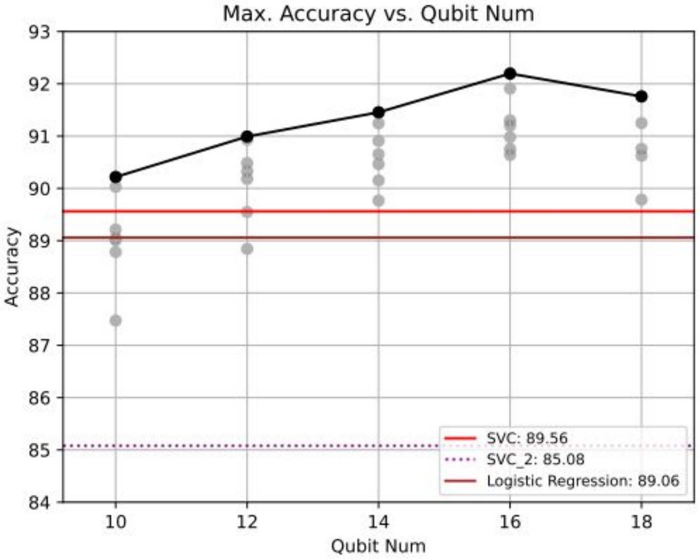
Training duration (800 epochs, single sQE architecture):

- 10-qubit, A100 : ~6.8 hours
- 10-qubit, L4: ~9.1 hours
- 16-qubit, A100 : ~43 hours
- 16-qubit, L4: ~56 hours
- 18-qubit, H100: ~90 hours

Quantum + AI fine-tuning: results on accuracy

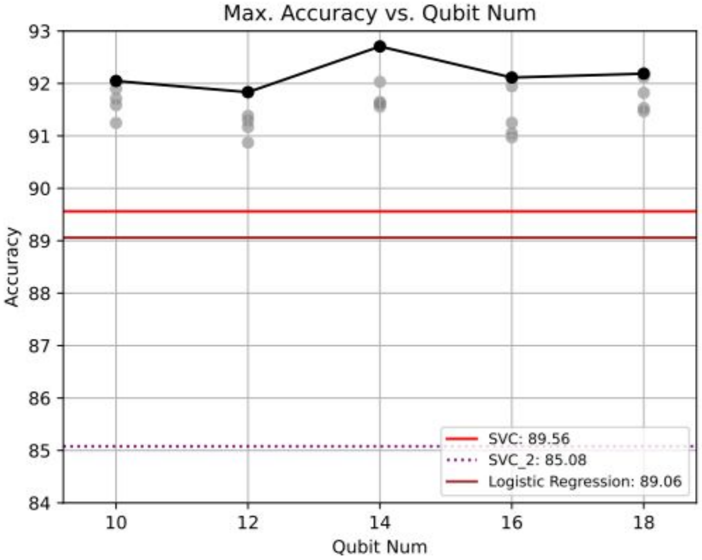
- Shot noise, No gate noise, # of epochs: 800, # of random seeds: 10, # of shots: 8192
- ReUpload #: 4, Layer # (PQC2): 2, Layer # (PQC1): 1, Connectivity: 1, Batch Size: 16, Final Linear Layer: Yes

Single Encoder



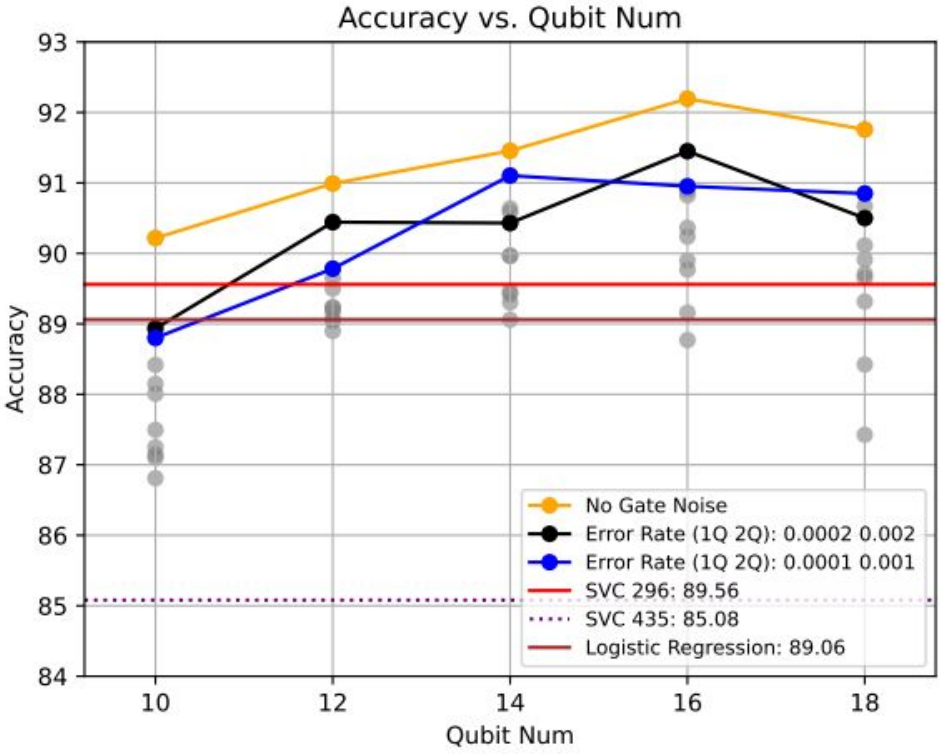
Learning Rate:0.0005, 0.001, 0.0015, 0.002, 0.0025, 0.003, 0.005
Learning Rate Decay: 1.0
L2 Weight Decay (Regularization): 0.0

Multi Encoder (2x)



Learning Rate: 0.001, 0.0015, 0.0025, 0.003, 0.005
Learning Rate Decay: 1.0, 0.99
L2 Weight Decay (Regularization): 0.0

Quantum + AI fine-tuning: results on gate- and shot-noise



IONQ

Conclusions

We demonstrated a hybrid quantum-classical architecture for LLM fine tuning showing:

- **Improved accuracy in low-data regimes**
- **Enhanced expressivity through quantum circuits**
- **Energy efficiency at scale**
- **Parameter-efficient fine-tuning**

Next steps

- **Quantum validation:** Run on QPU (Forte Enterprise). Measure energy consumption
- **Hyperparameter sweep:** Use Frontier! Engaged with AMD and Xanadu (Pennylane) at Spring Hackathon, ongoing development.
- **Base-model re-training:** gentle re-training of the base model.
- **New language use cases:** more classes; explore other tasks
- **New Foundation AI:** Chemistry, materials, biology, etc etc

IONQ