International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 04 April-2026 | Impact Factor: 3.5

PEFT Fine-Tuning with 1.58-Bit Quantization for a Quantum Computing Research Agent Chatbot: Architecture, Mathematical Foundations, and Practical Implementation

Rasikannan.L[1], Mohamed Israk.B[2], Vijay.K[3], Vinoth Kumar.G[4], Sabtharishi.G[5]
Department of Computer Science and Engineering
Government College of Engineering Srirangam, Trichy, Tamil Nadu, India

How to Cite this Article:
Rasikannan.L, , Israk.B, M., Vijay.K, , Kumar.G, V. & Sabtharishi.G, (2026). PEFT Fine-Tuning with 1.58-Bit Quantization for a Quantum Computing Research Agent Chatbot: Architecture, Mathematical Foundations, and Practical Implementation. International Journal of Creative and Open Research in Engineering and Management, 02(04). https://doi.org/10.55041/ijcope.v2i4.942

https://doi.org/10.55041/ijcope.v2i4.942

Abstract— Quantum computing represents one of the most technically dense frontiers in modern science, with specialized vocabulary spanning quantum gates, superposition, entanglement, variational quantum algorithms, and error-correction codes. Providing accurate, real-time research-grade assistance in this domain requires an LLM that is both highly capable and deployable on resource-constrained hardware. This paper presents the design and deployment of a Quantum Computing Research Agent Chatbot powered by PEFT fine-tuning of a 1.58-bit quantized LLM (BitNet b1.58). We extend the mathematical frameworks of Low-Rank Adaptation (LoRA), Adapter Layers, and Prefix Tuning to the quantum-domain fine-tuning context, incorporating a bespoke quantum-terminology corpus and a Retrieval-Augmented Generation (RAG) layer backed by a curated quantum literature index. We derive the rank–accuracy trade-off in the context of quantum NLP tasks, show that LoRA at rank r = 16 achieves an F1 score of 88.7 on quantum question-answering benchmarks while using only 0.13% of full model parameters, and demonstrate end-to-end inference at 4.2 tokens/second on a single NVIDIA T4 GPU. The proposed agent pipeline integrates with academic literature APIs (arXiv, Semantic Scholar) and supports multi-turn research dialogues, citation generation, and mathematical expression rendering.

Keywords— *Quantum Computing Chatbot, PEFT, BitNet 1.58-bit Quantization, Low-Rank Adaptation, Research Agent, Retrieval-Augmented Generation, Ternary Weight Matrices, Quantum NLP*

I. INTRODUCTION

Quantum computing has undergone a rapid transition from theoretical abstraction to noisy intermediate-scale quantum (NISQ) reality. With processors from IBM, Google, IonQ, and Quantinuum reaching 100+ qubit scales, the research community faces an information overload: hundreds of arXiv preprints per week, overlapping notational conventions, and highly interdisciplinary literature spanning physics, mathematics, and computer science. Researchers—especially students and interdisciplinary practitioners—require intelligent assistants capable of explaining concepts ranging from the Hadamard transform to the quantum approximate optimization algorithm (QAOA) with mathematical rigor and pedagogical clarity.

General-purpose LLMs such as GPT-4 and LLaMA-3 possess broad linguistic capability, but their quantum-domain accuracy is limited by insufficient representation of quantum-specific terminology and mathematical formalism in general pre-training corpora. Fine-tuning these models on quantum literature dramatically improves domain performance, yet full fine-tuning of a 7B-parameter model in FP16 requires over 70 GB of GPU memory—prohibitive for most academic labs.

Two complementary technologies bridge this gap. First, **BitNet b1.58 quantization** [Ma et al., 2024] compresses every weight to the ternary alphabet $W \in \{-1, 0, +1\}$, reducing storage by 8–10× versus FP16 while preserving model quality within 1–2% on standard benchmarks. Second, **Parameter-Efficient Fine-Tuning (PEFT)**—specifically LoRA, Adapter Layers, and Prefix Tuning—fine-tunes only a small trainable subspace (< 0.5% of parameters) while freezing the ternary backbone.

This paper makes the following contributions:

- A Quantum Computing Research Agent Chatbot architecture integrating a PEFT-fine-tuned 1.58-bit LLM with a RAG layer backed by 42,000 quantum literature documents.
- Extension of the quantization-aware PEFT mathematical framework to quantum-domain NLP tasks, including derivation of the rank–accuracy trade-off for quantum terminology density.
- An empirical evaluation on three quantum-domain benchmarks: QuALITY-QC, QuantumMath, and QCircuit-Explain.
- A practical deployment blueprint achieving 4.2 tokens/second on a T4 GPU with a 4.6 GB memory footprint.

II. BITNET 1.58-BIT QUANTIZATION: BACKGROUND

A. Ternary Weight Representation

Let $W \in \mathbb{R}^{m\times n}$ be a pre-trained weight matrix. BitNet b1.58 maps W to a ternary matrix $\tilde{W} \in \{-1, 0, +1\}^{m\times n}$ via the absmean quantizer:

$$\tilde{W} = RoundClip(W / \gamma, -1, +1) \qquad (1)$$

where $\gamma = (1/mn)\,\Sigma|W_{ij}|$ is the absolute-mean scaling factor. Activations are quantized to 8-bit integers via per-token scaling. The effective matrix–vector multiplication separates the discrete combinatorial structure $\tilde{W}$ from continuous scalar rescaling (γ, β), enabling efficient integer-arithmetic inference with 8–10× memory compression relative to FP16.

B. Spectral Properties of Quantum-Domain Ternary Weights

After domain-adaptive pre-training on quantum literature, weight matrices exhibit a **steeper singular value decay** than general-domain weights. The narrower distributional support of quantum-domain text (restricted vocabulary of gates, operators, and circuit primitives) concentrates signal in fewer principal components: the cumulative explained variance of the top-16 singular values rises to 85–90% (versus 78–84% for general-domain weights). This implies that even lower ranks $r \leq 8$ may suffice for quantum-domain fine-tuning, as confirmed by our ablation in Section V.

The SVD of the ternary matrix $\tilde{W} = U\Sigma V^T$ satisfies $\|\tilde{W}\|_F^2 = \Sigma_i \sigma_i^2 = nnz(\tilde{W})$, since every non-zero entry contributes exactly ±1 to the squared norm. The Moore-Penrose pseudoinverse $\tilde{W}^+ = V\Sigma^{-1}U^T$ and the column-space / null-space partition of $\mathbb{R}^m$ are central to the LoRA adapter design in Section III.

III. PEFT METHODS: MATHEMATICAL FOUNDATIONS

A. Low-Rank Adaptation (LoRA)

The adapted forward pass for a frozen ternary weight $\tilde{W}$ is:

$$h = (\tilde{W} + \Delta W)x = \tilde{W}x + BAx, \quad A \in \mathbb{R}^{r\times n}, \; B \in \mathbb{R}^{m\times r} \qquad (2)$$

For quantum-domain fine-tuning, the LoRA adapter captures domain-shift directions absent from the general pre-training distribution. The column-space decomposition $B = U_{\parallel} B_{\parallel} + U_{\perp} B_{\perp}$ (where $U_{\perp}$

spans directions orthogonal to Col($\tilde{W}$)) is critical: the component $B_\perp B_\perp^T A A^T$ adds directions *invisible* to $\tilde{W}$, encoding quantum-specific concepts (qubit register notation, quantum channel representations, stabilizer tableau algebra) that lie outside the column span of the general-domain ternary weights. The trainable parameter count $r(m+n) \ll mn$; for $d = 4096$ and $r = 16$ this is 0.13% of full parameters.

B. Gradient Flow

The gradient update equations for LoRA matrices in the quantum-domain setting are:

$$\nabla_A L = B^T \nabla_W L \cdot x^T \quad (3) \qquad \nabla_B L = \nabla_W L \cdot x \cdot A^T \quad (4)$$

Since $\tilde{W}$ is frozen, the Straight-Through Estimator for the absmean quantizer is not required during adapter training. The Riemannian gradient on the rank-r manifold M_r is approximated by the LoRA parameterization BA, preventing catastrophic forgetting of the ternary backbone's general-language capabilities while injecting quantum-domain knowledge.

C. Adapter Layers and Prefix Tuning

Adapter layers insert bottleneck modules after each transformer sub-layer:

$$h_{out} = h + f(h W_{down}) W_{up}, \quad W_{down} \in \mathbb{R}^{d \times k}, W_{up} \in \mathbb{R}^{k \times d} \quad (5)$$

For the quantum research agent, adapters at $k = 64$ are placed after feed-forward sub-layers, capturing domain-specific transformations (e.g., mapping natural-language quantum gate descriptions to their unitary matrix representations). Prefix tuning is applied to attention layers with $P = 20$ prefix tokens initialized from quantum-glossary embeddings, encoding persistent quantum context across all attention heads.

IV. QUANTUM RESEARCH AGENT: ARCHITECTURE AND OPTIMIZATION

A. System Overview

The Quantum Computing Research Agent Chatbot is a multi-component pipeline with the following major modules:

- PEFT-Fine-Tuned 1.58-Bit LLM Core – A 7B-parameter ternary LLM (BitNet b1.58 base) fine-tuned with LoRA (r = 16) on QuantumCorpus-2024.
- RAG Layer – A FAISS vector index over 42,000 quantum computing papers (arXiv cs.QC, quant-ph, 2015–2024) retrieved using a quantum-aware embedding model (SciBERT fine-tuned on quant-ph abstracts).
- Multi-Turn Dialogue Manager – Maintains sliding-window conversation history of 3,800 tokens with a DistilBERT compression model summarizing older exchanges.

- Mathematical Expression Renderer – Detects LaTeX expressions in LLM output and renders them via MathJax 3.0, with post-processing validation for unitarity of gate matrices and positive semidefiniteness of density matrices.
- Literature API Connector – Live queries to arXiv and Semantic Scholar APIs for paper retrieval, BibTeX citation generation, and abstract summarization.

B. QuantumCorpus-2024 Fine-Tuning Dataset

The fine-tuning corpus consists of 2.1M instruction-response pairs constructed from: (i) quantum computing Q&A extracted from Physics Stack Exchange and Quantum Computing Stack Exchange; (ii) textbook exercise solutions from Nielsen & Chuang and Preskill's lecture notes; (iii) 15,000 arXiv abstract-to-layman-summary pairs; and (iv) 8,500 quantum circuit description pairs generated as silver labels. The corpus covers quantum gates, quantum algorithms (Shor's, Grover's, QAOA, VQE), quantum error correction (surface codes, stabilizer formalism), quantum communication (BB84, E91), and quantum hardware (superconducting qubits, trapped ions, photonics).

C. Joint Optimization Objective

The quantum research agent is trained to minimize:

$$\min_{A,B} L_{QC}(f(Q(W) + BA; x_q), y_q) \quad s.t. \quad \text{rank}(BA) \leq r \qquad (6)$$

where x_q is a quantum-domain user query and y_q is the ground-truth response. L_QC combines token-level cross-entropy with a domain-specific symbol-consistency penalty $\lambda \cdot L_{sym}$ ($\lambda = 0.1$) that penalizes confusing $|0\rangle$ and $|1\rangle$ state labels or producing non-unitary gate matrices. The quantization gap is bounded by:

$$L(\theta^*_{PEFT}) - L(\theta^*_{full}) \leq C_{QC} \cdot \gamma^2 \cdot r \qquad (7)$$

The constant C_QC is *smaller* for quantum-domain text than for general text due to steeper singular value decay, implying that a given accuracy threshold is reached at a lower rank r. Our ablation confirms that r = 8 already achieves 85.4 F1 on quantum QA, while r = 16 reaches 88.7 F1 with only a modest increase in memory (4.3 GB → 4.6 GB).

V. EXPERIMENTAL RESULTS

A. Experimental Setup

We evaluate the PEFT-fine-tuned 1.58-bit Quantum Research Agent on three benchmarks: (i) **QuALITY-QC** – 500 multiple-choice quantum computing QA pairs; (ii) **QuantumMath** – 300 formula-grounded reasoning problems; and (iii) **QCircuit-Explain** – 200 quantum circuit description tasks. PEFT method variants are compared against the FP16 full fine-tuning baseline. All PEFT experiments use a single NVIDIA T4 16 GB GPU; the FP16 baseline uses an A100 40 GB GPU.

B. Performance Results

Table I: Performance of PEFT Methods on 1.58-Bit Quantum Computing Research Agent

Method	Params (M)	QuALITY-QC F1	QuantumMath Acc	QCircuit BLEU	GPU Mem (GB)
Full FT (FP16)	7000	91.2	88.3	38.1	38.0
LoRA r=4 (1.58b)	2.4	83.6	79.4	33.2	4.1
LoRA r=8 (1.58b)	4.7	85.4	82.1	35.0	4.3
LoRA r=16 (1.58b)	9.4	88.7	85.6	37.2	4.6
Adapter (1.58b)	6.1	86.3	82.9	35.8	4.5

Method	Params (M)	QuALITY-QC F1	QuantumMath Acc	QCircuit BLEU	GPU Mem (GB)
Prefix (1.58b)	1.8	81.5	77.2	31.9	4.0

Table I shows that LoRA r=16 on the 1.58-bit quantum research agent achieves **88.7 F1 on QuALITY-QC**, within 2.5 points of FP16 full fine-tuning, while requiring only **4.6 GB of GPU memory**—comfortably within T4 capacity. Prefix tuning shows the smallest parameter footprint but the largest accuracy drop, consistent with the rank–accuracy bound (Eq. 7). The symbol-consistency penalty L_sym reduces quantum-specific hallucinations (incorrect gate matrix entries, swapped state labels) by 34% compared to standard cross-entropy training.

C. Inference Speed and Deployment Metrics

End-to-end inference for a typical quantum research query ("Explain the variational quantum eigensolver and its convergence guarantees") proceeds at **4.2 tokens/second** on a T4 GPU with batch size 1. The RAG retrieval step adds a mean latency of 210 ms per query (FAISS nearest-neighbor search over 42,000 documents). Total first-token latency is 1.1 seconds.

VI. SYSTEM DESIGN: CHATBOT INTERFACE AND AGENT CAPABILITIES

A. Multi-Turn Research Dialogue

The dialogue manager maintains a sliding-window conversation history of up to 3,800 tokens, reserving 296 tokens for the retrieved context snippet. When the history exceeds the window, the oldest exchanges are summarized by a lightweight DistilBERT compression model fine-tuned on dialogue summarization, preserving key concepts (qubit counts, algorithm names, prior clarifications) across long research sessions.

B. Citation Generation and Literature API Integration

When the quantum research agent references a paper, it triggers a Semantic Scholar API call to retrieve canonical citation metadata, formatting BibTeX entries inline and supporting IEEE, ACM, and APA citation styles on request. For arXiv preprints, the agent additionally provides a hyperlinked abstract and a "Similar Papers" sidebar retrieved via the arXiv Semantic Search API.

C. Mathematical Expression Handling

Quantum computing responses frequently contain matrix expressions (e.g., Hadamard gate $H = (1/\sqrt{2})[[1,1],[1,-1]]$), Dirac notation ($|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$), and Pauli operator algebra ($\sigma_x, \sigma_y, \sigma_z$). The LLM is prompted via LoRA prefix tokens to output mathematically correct LaTeX, which is rendered client-side via MathJax 3.0. Post-processing validation checks that density matrices are positive semidefinite and gate matrices are unitary, flagging inconsistencies before display.

VII. CONCLUSION

This paper has presented a Quantum Computing Research Agent Chatbot that applies PEFT fine-tuning of 1.58-bit quantized LLMs to the highly specialized domain of quantum computing. Building directly on the mathematical framework of BitNet b1.58 quantization and LoRA fine-tuning, we demonstrated that the steeper singular value decay of quantum-domain weight matrices enables effective adaptation at ranks as low as r = 8, with r = 16 achieving 88.7 F1 on quantum QA—within 2.5 points of the full FP16 baseline while consuming only 4.6 GB of GPU memory. The system's integration of RAG, live literature API connectivity, mathematical expression rendering, and multi-turn dialogue management positions it as a practical research assistant for the quantum computing community. Future work will explore adaptive rank selection per transformer layer using quantum-domain matrix coherence criteria, and extend the framework to incorporate structured knowledge graphs of quantum algorithm taxonomies for improved multi-hop reasoning.

ACKNOWLEDGMENT

The authors thank the Department of Computer Science and Engineering, Government College of Engineering Srirangam, *Trichy* for computational and institutional support. The authors acknowledge the open-source contributions of the Hugging Face PEFT library and the BitNet community for foundational tooling, and the Quantum Computing Stack Exchange community for publicly available Q&A data used in QuantumCorpus-2024.

REFERENCES

[1] S. Ma et al., "The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits," arXiv:2402.17764, 2024.

[2] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," Proc. ICLR, 2022.

[3] N. Houlsby et al., "Parameter-Efficient Transfer Learning for NLP," Proc. ICML, 2019.

[4] X. L. Li and P. Liang, "Prefix-Tuning: Optimizing Continuous Prompts for Generation," Proc. ACL, 2021.

[5] T. Dettmers et al., "QLoRA: Efficient Finetuning of Quantized LLMs," Proc. NeurIPS, 2023.

[6] A. Aghajanyan et al., "Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning," Proc. ACL, 2021.

[7] M. A. Nielsen and I. L. Chuang, Quantum Computation and Quantum Information, Cambridge University Press, 2000.

[8] J. Preskill, "Quantum Computing in the NISQ Era and Beyond," Quantum, vol. 2, p. 79, 2018.

[9] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Proc. NeurIPS, 2020.

[10] G. H. Golub and C. F. Van Loan, Matrix Computations, 4th ed., Johns Hopkins University Press, 2013.

[11] Y. Bengio et al., "Estimating or Propagating Gradients Through Stochastic Neurons," arXiv:1308.3432, 2013.

[12] Z. Zhang et al., "Quantization-Aware Training for Natural Language Understanding," Proc. EMNLP, 2022.