

# QuantLayer: A Hybrid Fine-Tuned GraphRAG Architecture for Zero Trust Anomaly Detection in Industrial Control Systems

M. Afraz Ahmad

Senior Full Stack AI Engineer, BData Solutions

[afraz@bdata.ca](mailto:afraz@bdata.ca)

*Target Audience: Academic Research Groups & Enterprise Security Stakeholders*

January 20, 2026

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>System Architecture</b>	<b>2</b>
2.1	The Ingestion Layer (Event Backbone) . . . . .	4
2.2	The Hybrid Memory Layer . . . . .	4
2.3	Reasoning Engine . . . . .	4
<b>3</b>	<b>Methodology</b>	<b>4</b>
3.1	Fine-Tuning Strategy (QLoRA) . . . . .	4
3.2	GraphRAG: Overcoming Context Blindness . . . . .	5
3.3	Mathematical Formalization . . . . .	5
3.3.1	QLoRA Weight Adaptation . . . . .	5
3.3.2	Graph-Aided Retrieval Score . . . . .	5
3.4	Planning For Evaluation Framework . . . . .	5
<b>4</b>	<b>Scalability and Engineering</b>	<b>5</b>
<b>5</b>	<b>Future Research Directions</b>	<b>6</b>
<b>6</b>	<b>Conclusion</b>	<b>6</b>

## Abstract

Traditional anomaly detection used in Industrial Control Systems (ICS) excessively depends on static signatures, making them unable to detect complex or sophisticated multi-vector attacks. This paper presents **QuantLayer**, a new architecture of Zero Trust which is a combination of Large Language Models (LLMs) and Graph Retrieval-Augmented Generation (GraphRAG). By combining the probabilistic nature of a fine-tuned model and the deterministic nature of a Knowledge Graph (topological awareness), QuantLayer achieves ‘context-aware’ threat detection. For high-velocity ingestion, the system uses Apache Kafka, temporal coherency, and a hybrid retrieval engine to limit false positives in resource-constrained OT environments (utilizing TimescaleDB).

## 1 Introduction

The convergence of IT and OT (Operational Technology) networks has exposed critical infrastructure to Advanced Persistent Threats (APTs). While modern AI models have much potential in the field of anomaly detection, they present two major challenges in the industrial field:

- **Context Blindness:** Standard LLMs have a tendency to work on log text strings without considering context. They do not know the physical dependencies of the sensor (e.g., LIT-101) and the actuator (e.g., MV-101), thus resulting in high false positive rates.
- **Resource & Privacy Constraints:** It is computationally infeasible to run large-scale commercial models on edge gateways, and it presents an unacceptable risk to data sovereignty.

QuantLayer overcomes these disadvantages by proposing an **Agentic AI Architecture** localized at the edge. It does not just classify logs but is able to “reason” about system states by correlating real-time telemetry with a static knowledge graph of the physical topology of the plant.

## 2 System Architecture

QuantLayer architecture leverages a Microservices Pattern decoupled for high throughput inference (10k events/sec). The system consists of 3 main layers: Ingestion, Hybrid Memory, and Reasoning.

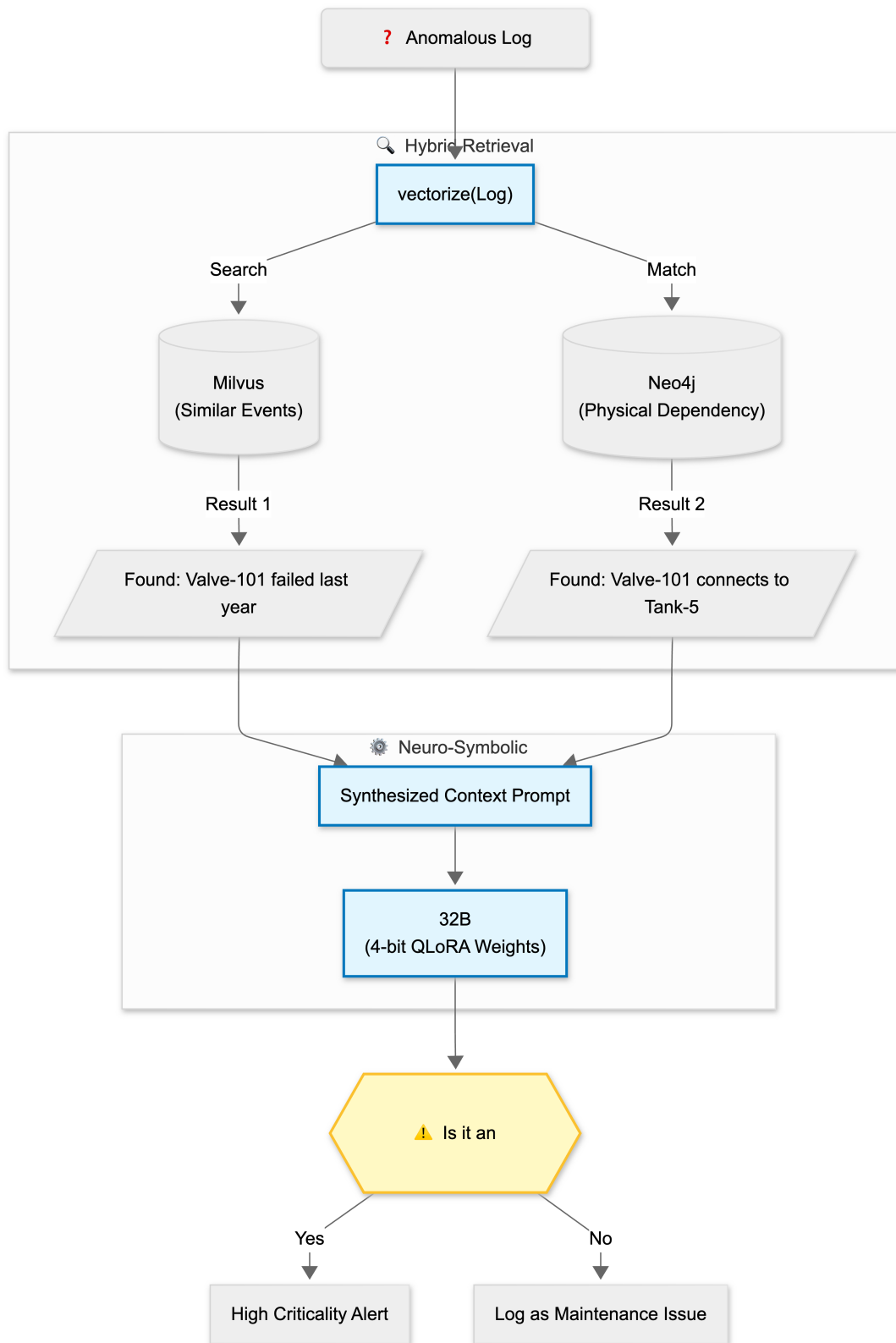


Figure 1: High Level System Architecture - showing the flow of data from the OT Edge to the Hybrid Memory Layer before interpretation.

## 2.1 The Ingestion Layer (Event Backbone)

Massive volumes of sensor telemetry are generated in OT environments. QuantLayer uses **Apache Kafka** as the centralized event backbone for decoupling data producers (PLCs/SCADA) from the analytical engine. This guarantees zero data loss during traffic spikes, as well as the ability to replay attack sequences to perform forensic analysis.

## 2.2 The Hybrid Memory Layer

To reduce hallucination risk in LLMs, QuantLayer uses a split-memory architecture:

- **Temporal Memory (TimescaleDB):** Represents raw time-series sensor data to calculate sliding statistical windows (e.g., moving averages).
- **Topological Memory (Neo4j):** Stores the “Physical Graph” of the facility representing the interconnectivity of sensors, PLCs, and physical assets (e.g., Purdue Model Level 1 devices).
- **Semantic Memory (Milvus):** Stores the vectorized representation of historical attack signatures and known anomaly descriptions.

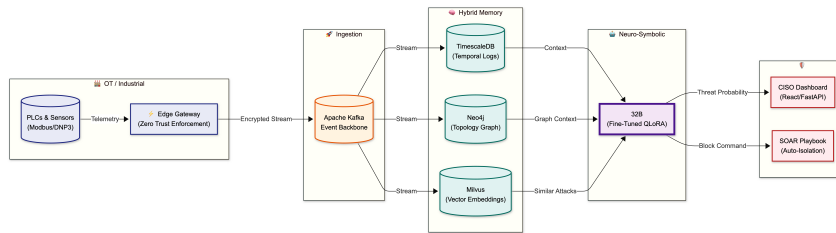


Figure 2: Hybrid Retrieval Logic. The model combines the context of Vector Search (Milvus) and Graph Search (Neo4j).

## 2.3 Reasoning Engine

The core of the system is a customized reasoning model. Unlike generic chat models, this model is a “Reasoning Agent” that follows Chain-of-Thought (CoT) logic to determine if a sensor deviation is a benign operational shift or malicious actuation.

# 3 Methodology

## 3.1 Fine-Tuning Strategy (QLoRA)

In order to make the model comprehend the heterogeneous syntax of OT protocols (Modbus/DNP3) and the physics of the industrial process, we used **QLoRA** (Quantized Low-Rank Adaptation).

- **Optimization:** The model weights are quantized to 4-bit NormalFloat (NF4) precision, allowing deployment onto consumer-grade GPUs (e.g., Nvidia A6000) whilst performing computation in BrainFloat-16 (BF16).
- **Dataset:** The model has been fine-tuned from a curated combination of **SWaT** (Secure Water Treatment) and **WaDi** (Water Distribution) datasets. This combination of benign operational logs and labeled multi-point attack vectors provides balance to the model.

### 3.2 GraphRAG: Overcoming Context Blindness

Standard RAG retrieves documents based solely on semantic similarity. In OT, this is insufficient. For example, high pressure is anomalous *only* when the outflow valve is closed. QuantLayer implements **GraphRAG**:

- **Vector Retrieval:** Milvus detects semantically similar historical errors.
- **Graph Traversal:** Neo4j looks for downstream impact. If Sensor A is aberrant, the agent consults the graph to get the state of Valve B connected to it.
- **Context Injection:** The LLM deals with the prompt using Context Augmented Historical Data (Vector) as well as physical reality (Graph).

### 3.3 Mathematical Formalization

We define the anomaly detection function  $f_\theta(x)$  not as a simple classification, but as a conditional probability distribution on system states, dependent on retrieved topological context.

#### 3.3.1 QLoRA Weight Adaptation

To respect edge resource limitations, we freeze the pre-trained weights  $W_0 \in \mathbb{R}^{d \times k}$  and inject trainable low-rank decomposition matrices  $A$  and  $B$ . The forward pass of the hidden layer  $h$  is defined as:

$$h = W_0x + \Delta Wx = W_0x + BAx \quad (1)$$

#### 3.3.2 Graph-Aided Retrieval Score

Standard RAG relies on Cosine similarity  $S_{cos}$ . QuantLayer introduces a topological relevance score. For a candidate node  $n_i$  and a query log  $q$ , the final relevance score  $R(n_i)$  is:

$$R(n_i) = \alpha(\vec{q} \cdot \vec{n}_i) + (1 - \alpha) \sum_{j \in N(n_i)} \frac{1}{\deg(n_j)} \quad (2)$$

### 3.4 Planning For Evaluation Framework

We will conduct an experimental study using the **SWaT** dataset. Performance will be measured against a baseline Standard RAG (Vector only) approach:

- **Precision/Recall:** To measure the reduction of false positives.
- **Mean Time to Detection (MTTD):** To assess the latency impact of dual-step graph traversal.
- **Hallucination Rate:** Using a ‘‘Factuality Score’’ by manually checking if the generated root cause matches ground truth attack labels.

## 4 Scalability and Engineering

The system is based on a Microservices Architecture implemented on **FastAPI** (Python) and **Docker**, guaranteeing modular scalability.

- **High Throughput:** The Kafka-based pipeline supports high-velocity telemetry (10k+ events/second).

- **Edge Optimization:** Inference is optimized using 4-bit quantization, ensuring the system runs efficiently on-site with minimal latency.
- **Data Sovereignty:** Adheres to strict requirements (GDPR, NERC CIP) ensuring OT data never leaves the facility.

## 5 Future Research Directions

- **Adversarial Robustness:** Studying “Agentic” models against adversarial noise injection (EGSM attacks).
- **Federated Learning:** Exchanging threat intelligence gradients across facilities without sharing raw proprietary topology data.
- **Explainability (XAI):** Visualizing “reasoning steps” back to physical graph nodes for operator trust.

## 6 Conclusion

QuantLayer represents a shift from static detection rules to dynamic, context-aware reasoning. By offering the probabilistic capability of Fine-Tuned LLMs combined with the deterministic structure of Knowledge Graphs, we propose a robust architecture for the next generation of Critical Infrastructure protection.

## References

- [1] Lewis, P. et al. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. NeurIPS.
- [2] Edge, D. et al. (2024). *From Local to Global: A Graph RAG Approach to Query-Focused Summarization*. Microsoft Research.
- [3] Hu, E. J. et al. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. ICLR.
- [4] Goh, J. et al. (2016). *A Dataset to Support Research in the Design of Secure Water Treatment Systems (SWaT)*. Critical Infrastructure Protection.