Politecnico di Torino

北京航空航天大學
BEIHANG UNIVERSITY

Master of Science in Industrial Production and Technological Innovation Engineering

Digital Twin International Research Center

Cognitive Digital Twins for Quality Governance in Smart Factories: An LLM-Based Architectural Framework

Candidate:
Mirco Hu

Supervisors:
Prof. Alessandro Simeone
Prof. Li Yi

Co-*Supervisor:*
Prof. Domenico A. F. Maisano

Academic Year 2025/2026

There are people without whom this journey would have taken a very different shape, and for whom no acknowledgement feels sufficient.

I begin with my thesis supervisors, Professor Alessandro Simeone and Professor Yi Li. You gave me something I did not expect: the opportunity to write this thesis in Beijing, in China, in the country that is half of who I am. That decision changed everything. Being in Beijing was not only an academic experience; it was a rediscovery of my origins, of a culture that had always lived inside me but that I had never fully encountered. For that gift, and for the trust you placed in me, I am profoundly grateful.

The person I want to thank most is Pia, my girlfriend, my balance, the most unexpected and beautiful thing that Barcelona brought me. I went there believing I was simply going to study abroad. I did not imagine I would find someone who would quietly become one of the most important parts of my life. You fill the spaces I did not even know I had. You give me what I had never found before, and you do it with a naturalness that still amazes me. We move in harmony, in the chaos of daily life, through a thousand adventures between China, Spain and Italy. Thank you for your patience, for the way you have helped me grow and mature, for the moments of joy that were simple because you were there, and for always being there for me, not only in the difficult moments. I did not think a person like you existed. I am glad I was wrong.

最重要的是，我要向我的父母献上最深的感谢。这篇论文，我整个求学历程，都献给你们。没有你们，我什么都不是。感谢你们传授给我人生中最重要的价值观：谦逊、永不放弃的精神、随遇而安的能力。感谢你们为我所做的一切牺牲，那些你们从未开口提起、却我深深铭记于心的牺牲。你们是我所有努力的根基，是我前行的力量。

I want to thank Luigi. We started this adventure together two years ago, and we are bringing it to the end together. I hope our paths will keep crossing in the future, who knows where life will take us next. Through you I discovered that patience is not passivity, and that the best friendships are built in the ordinary moments, the ones you do not notice until later. I also discovered an extraordinary friend, and I am glad our paths have stayed intertwined across different cities and chapters of life.

I never expected to find a friend like Raffaella, someone so warm, so thoughtful, so genuinely caring. You appeared and became a constant throughout this whole adventure, and I am grateful for everything that means. Thank you for always being there, even when I did not ask.

A mention goes to Davide and Domenico. I wanted to acknowledge the time we spent together during this adventure. It was a genuine pleasure, and I am glad our paths crossed.

I cannot name everyone, there are too many, but know that you are all in my thoughts. I want to acknowledge the people I met during my year in Barcelona, my classmates and course companions, who kept travelling this journey alongside me. You came from everywhere and brought everything with you. Through you I grew, matured, and discovered that there is no obstacle that a change of perspective and good company cannot make smaller. You made me understand that the world is much more open than it sometimes appears, and I carry that with me.

And even though our time together was short, to the guys who were there alongside me in China at Beihang University: I discovered truly wonderful people in you, and that made everything even better.

To the people of Carrer de Tiradors: thank you for the home we made together. To my friends in Florence, ϑ, you know who you are. To my Kenyan family from Thika: thank you for being there, and for keeping this bond alive long after we left. And to NRT, my family of a lifetime, the thread that runs through all of it, thank you for simply always being there.

This is not just the end of a thesis. It is the end of a university path that gave me people I did not know I needed. I am grateful for all of it, and for everyone who walked part of the way with me.

Thanks. Grazie. Gracias. Merci. 谢谢.

Abstract

The use of IoT-enabled Digital Twins with Strategic Quality Management in the context of Smart Factories has reached considerable technical maturity. Nevertheless, a cognitive gap has persisted: the high fidelity, physics based intelligent insight from Digital Twins has been unavailable to operational staff without access to complex software; and at the same time, there is no specific architecture to link Digital Twin process representation and the semantic reasoning capability of Large Language Models (LLM's).

We can highlight five main research gaps: an interface and interpretability bottleneck, gaps in cyber-physical system orchestration and cyber-physical process modelling, a need for semantic reasoning layer and improved knowledge management, and gap in standardization.

In response, this thesis proposes the LLM-DT-SQM framework, a five-layer closed-loop hybrid architecture integrating IoT-enabled Digital Twins with Large Language Model capabilities for Strategic Quality Management in Smart Factory environments. The Input Layer for IIoT and enterprise data acquisition; a Digital Twin and Knowledge Layer housing a four-capability cyber-physical process model and a causal knowledge base; a Generative AI Layer implementing a six-stage cognitive pipeline comprising natural language understanding, Retrieval-Augmented Generation context assembly, Knowledge Graph-grounded causal reasoning, API function calling, insight generation, and automated validation; a Human Interaction Layer featuring two formal governance gates and human approval authority; and a Data Management Layer providing immutable audit trail management and Bayesian closed-loop model learning. All layers are supported by a transversal API and Integration Backbone providing unified communication, security, and protocol adaptation.

The framework is evaluated through a simulation-based industrial case study of a precision CNC machining line for automotive powertrain cylinder heads, adjusted by two industrial datasets. An evaluation across four representative quality scenarios yields an 82% reduction in CAPA cycle time from 47 to 8.4 minutes, 86.7% root-cause analysis precision for known failure modes, and full IATF 16949:2016 and ISO 9001:2015 compliance traceability. The principal limitation is Knowledge Graph coverage dependency: precision declines for failure modes absent from the FMEA ontology, confirming that LLM reasoning amplifies available knowledge but cannot create new expertise in a domain-specific setting.

The thesis concludes by addressing the three research questions, acknowledging limitations, and outlining future research directions including multi-agent LLM orchestration, on-premises deployment, and cross-factory validation studies. The proposed framework constitutes the first explicit architecture for LLM-enhanced Digital Twins dedicated to Strategic Quality Management, offering a pathway toward autonomous, conversational, and human-centric Quality 4.0 systems.

Table of Contents

1 Introduction 17
1.1 Premise: Industrial Context 17
1.2 Problem Statement and Research Motivation 17
1.3 Objectives and Research Questions 19
1.4 Literature Research Methodology 20
1.4.1 Phase 1 – Targeted Review: IoT–DT–SQM 22
1.4.2 Phase 2 – Expanded Review: LLM/API in Broader Industrial Settings 25
1.5 Summary of Identified Research Gaps 26
1.6 Thesis Structure 27
2 Literature Review 28
2.1 Introduction: The Convergence of Cyber-Physical Systems and SQM 28
2.1.1 Technical Foundations of IoT-Enabled Digital Twins 29
2.2 Predictive Maintenance as Strategic Quality Tool 30
2.2.1 Big Data Analytics 32
2.3 Current Industrial Applications of DT-IoT for SQM 33
2.4 Emerging Challenges and Research Gaps 34
2.4.1 The Interface and Interpretability Bottleneck 34
2.4.2 Fragmentation in Cyber-Physical Orchestration 35
2.5 Bridging the Gap: LLM-Enhanced DTs for SQM in Smart Factories 36
2.5.1 Large Language Models (LLMs) as Orchestrators of Industrial Knowledge 38
2.5.2 From Static Knowledge to Real-Time Action: The Role of Function Calling 39
2.5.3 API-Centric Design: Connecting the "Brain" (LLM) to the "Senses" (IoT/DT) 40
2.5.4 Architectural Synergy: Integrating LLMs into the DT Framework 41
2.6 LLM-Driven Strategic Quality Management 42
2.6.1 Natural Language Querying for Real-Time Quality Audits 42
2.6.2 Automated Root Cause Analysis (RCA) and Corrective Action Reports 43
2.7 State-of-the-Art: Early Applications of LLM-Enhanced DTs 44
2.8 Security, Ethics, and Latency Constraints 46
3 LLM-DT-SQM Framework 49
3.1 Purpose and Role of the LLM-DT Hybrid Architecture 49
3.2 Conceptual Architecture: LLM-DT-SQM Framework 50
3.2.1 Justification of Key Design Choices 52
3.2.2 Framework Overview 52
3.3 API & Integration Backbone 55
3.3.1 Communication Pattern: Unidirectional and Bidirectional Connections 56
3.4 Input Layer 57
3.4.1 Physical - Shop Floor 58
3.4.2 Enterprise Data 60

3.5 Digital Twin & Knowledge Layer 60
3.5.1 Digital Twin 61
3.5.2 Knowledge 63
3.5.3 Process State & KPI Export 63
3.6 Generative AI Layer 64
3.6.1 LLM Cognitive Layer: Six-Stage Processing Pipeline 64
3.6.2 Prompt Engineering and Context Window Design 67
3.7 Human Interaction Layer 68
3.7.1 SQM Application 69
3.7.2 Human Review & Approval 69
3.7.3 Final Approval 71
3.8 Data Management Layer 72
3.8.1 Database Storage and Audit Trail & Logging 72
3.8.2 Deployment & Feedback 73
3.9 Functional Modules 74
3.9.1 Natural Language Quality Audit Engine 74
3.9.2 Automated RCA & CAPA Generator 74
3.9.3 PdM-Quality Orchestrator 74
3.9.4 Compliance & Traceability Checker 74
3.9.5 Conversational Decision Support 75
3.10 Methodological Cycle 75
3.11 Safety, Ethics, and Governance 77
4 Case Study 78
4.1 Introduction to the Case Study 78
4.1.1 Case Study Selection Rationale 81
4.2 Objective and Scope 82
4.2.1 Simulation Parameter Derivation 84
4.2.2 Dataset Utilization: Parameter Extraction 86
4.3 Framework Implementation 89
4.3.1 Input Layer 89
4.3.2 Digital Twin & Knowledge Layer 90
4.3.3 Generative AI Layer 92
4.3.4 Human Interaction Layer 93
4.3.5 Data Management Layer 93
4.4 Implementation Workflow 94
4.5 Discussion 97
5 Results and Evaluation 98
5.1 Experimental Design and Evaluation Scenarios 98
5.1.1 Scenario S1: Baseline Operation 99

5.1.2 Scenario S2: Known Quality Anomaly (Tool Wear Drift) 99
5.1.3 Scenario S3: Multi-Variate Anomaly 99
5.1.4 Scenario S4: Unforeseen Failure Mode 100
5.2 Performance Indicators 100
5.3 Scenario-Based Results 103
5.3.1 S1: Baseline Operation 103
5.3.2 S2: Known Quality Anomaly 104
5.3.3 S3: Multi-Variate Anomaly 105
5.3.4 S4: Edge Case 105
5.4 Comparative Analysis 106
5.4.1 Framework Layer Verification Summary 108
5.5 Limitations and Critical Discussion 109
6 Conclusions and Future Developments 110
6.1 Summary of Research Contributions 110
6.2 Responses to Research Questions 110
6.3 Theoretical and Practical Implications 111
6.4 Limitations 112
6.5 Future Research Directions 113
6.6 Concluding Remarks 113
7 References 114

List of Figures

Figure 1 - Steps of the conducted study 20
Figure 2 - Flowchart of the Systematic Literature Review 24
Figure 3 - Interaction of real-time data using cloud database and digital twins 28
Figure 4 - High-level of edge, fog, and cloud layers 29
Figure 5 - LLM-DT-SQM framework 53
Figure 6 - Aluminum Alloy Cylinder Head for a Four-Cylinder Passenger Vehicle Engine 79
Figure 7 - DMG Mori NHX 5000 79
Figure 8 - Taylor-extended flank-wear progression (n = 0.35, NUAA dataset) 88

List of Tables

Table 1 - Layered Big Data Analytics Architecture for IoT-DT Quality Intelligence 32
Table 2 - Cross-Sector Industrial Applications of DT-IoT for SQM 33
Table 3 - Research Gaps in DT-IoT SQM Systems 36
Table 4 - API Function Types and their SQM Applications 40
Table 5 - LLM Capabilities and their Evidence-Based SQM Applications 42
Table 6 - State-of-the-Art LLM-Enhanced Digital Twin Applications 45
Table 7 - Security, Ethics, and Operational Constraints 48
Table 8 - Explicit mapping of the five research gaps to the framework components 51
Table 9 - Functional summary of the five layers and the API & Integration Backbone. 54
Table 10 - Backbone connection pattern per layer: unidirectional vs. bidirectional 57
Table 11 - Industrial Communication Protocol Selection for the Physical sub-layer 60
Table 12 - Digital Twin four-capability architecture 62
Table 13 - LLM API Function Catalog 66
Table 14 - Context window assembly 67
Table 15 - Key parameters and scope definition 84
Table 16 - Simulation parameter provenance 86
Table 17 - Workflow Timestamp Derivation 96
Table 18 - Overview of evaluation scenarios S1–S4 98
Table 19 - Performance indicators P1–P12 aligned with Chapter 3 targets. 101
Table 20 - Indicator-to-scenario applicability mapping 102
Table 21 - Scenario S1 (Baseline Operation) results. 103
Table 22 - Scenario S2 (Known Anomaly, Tool Wear Drift) results across 15 events. 104
Table 23 - Scenario S3 (Complex Multi-Variate Anomaly) results across 8 events 105
Table 24 - Scenario S4 (Edge Case, Catastrophic Insert Micro-Fracture) results. 106
Table 25 - Comparison LLM-DT-SQM framework vs pre-deployment SQM. 108

List of Abbreviations

AI – *Artificial Intelligence*
API – *Application Programming Interface*
BDA – *Big Data Analytics*
CAPA – *Corrective and Preventive Action*
CNC – *Computer Numerical Control*
CPPS – *Cyber-Physical Production Systems*
CPS – *Cyber-Physical Systems*
DES – *Discrete Event Simulation*
DRL – *Deep Reinforcement Learning*
DT – *Digital Twin*
ERP – *Enterprise Resource Planning*
FMEA – *Failure Mode and Effects Analysis*
GDPR – *General Data Protection Regulation*
H&S – *Health and Safety*
HTTP – *HyperText Transfer Protocol*
IATF – *International Automotive Task Force*
IEC – *International Electrotechnical Commission*
IIoT – *Industrial Internet of Things*
IoT – *Internet of Things*
ISO – *International Organization for Standardization*
JSON – *JavaScript Object Notation*
JSON-LD – *JSON for Linking Data*
JWT – *JSON Web Token*
KPI – *Key Performance Indicator*
LLM – *Large Language Model*
LSTM – *Long Short-Term Memory*
MES – *Manufacturing Execution System*
ML – *Machine Learning*
MQTT – *Message Queuing Telemetry Transport*
NCR – *Non-Conformance Report*
NLG – *Natural Language Generation*
NLU – *Natural Language Understanding*
NTP – *Network Time Protocol*
OAuth2 – *Open Authorization 2.0*
OEE – *Overall Equipment Effectiveness*
OPC-UA – *OPC Unified Architecture*
OWL – *Web Ontology Language*
PdM – *Predictive Maintenance*
PLC – *Programmable Logic Controller*
PLM – *Product Lifecycle Management*
PROFINET – *Process Field Net*
PTP – *Precision Time Protocol*
QMS – *Quality Management System*
QoS – *Quality of Service*
RAG – *Retrieval-Augmented Generation*
RCA – *Root Cause Analysis*
RDF – *Resource Description Framework*
REST – *Representational State Transfer*
RPN – *Risk Priority Number*
SCADA – *Supervisory Control and Data Acquisition*
SLR – *Systematic Literature Review*
SM – *Smart Factory*
SPARQL – *SPARQL Protocol and RDF Query Language*
SPC – *Statistical Process Control*
SQM – *Strategic Quality Management*
TQM – *Total Quality Management*
UNS – *Unified Namespace*
WORM – *Write-Once Read-Many*
XML – *eXtensible Markup Language*

1 Introduction

1.1 Premise: Industrial Context

The advent of Industry 4.0 has drove the manufacturing industries towards the development of smart and interconnect systems (Obasi et al., 2026; Sun et al., 2025), where the integration of the Internet of Things (IoT) and Digital Twin (DT) technologies has become pivotal for advancing Strategic Quality Management (SQM) within Smart Factories (Christou et al., 2022). This integration allows, on the other hand, to achieve a paradigm shift in how we conduct Quality Control, by facilitating real-time data collection, full process simulation, and intelligent predictions (Sokolov et al., 2023). The IoT infrastructure, consisting of various sensor networks and communication protocols, plays a crucial role, offering the unique ability to allow continuous synchronized communication between a physical system and its digital twin, a function necessary for micro-level quality monitoring through product life cycles (Bongomin et al., 2025; Tiwari et al., 2025; Tonelli et al., 2026).

This synergistic integration makes IoT-enabled Digital Twins as the primary enablers of Quality 4.0, a paradigmatic evolution of Strategic Quality Management from reactive inspection toward data-driven, predictive, and continuous improvement governance (Attia H., 2025).

Within Smart Factories, cyber-physical systems, IoT, and Digital Twins jointly create a closed-loop environment where quality-related decisions can be simulated, evaluated, and implemented in real time. These factories serve as the operational context where IoT-enabled Digital Twins translate strategic quality objectives into concrete actions on machines, processes, and supply chains enabling resilient, energy-efficient, and high-quality production systems (Friederich et al., 2022).

Despite these advances, the integration of LLM and API architectures as a semantic reasoning layer within IoT-DT-SQM systems remains unexplored in the academic literature, defining the central research gap addressed by this thesis.

1.2 Problem Statement and Research Motivation

While *Chapter 2* presents a detailed systematic literature review, this section introduces the core problem and the main challenges addressed by this thesis. Despite the remarkable progress achieved in IoT-enabled digital twins, smart factory implementation and Strategic Quality Management (SQM), several open challenges persist at both technological and managerial levels (Tonelli et al., 2026).

From a technological standpoint, many industrial implementations treat IoT, Digital Twin, and analytics as architecturally independent systems rather than as components of an

integrated, closed-loop cyber-physical quality governance framework. IoT infrastructures are widely deployed for real-time data acquisition from production assets, yet the translation of raw sensor data into actionable quality insights remains inconsistent and often manual (Bongomin et al., 2025; Onaji et al., 2022). Digital twins are frequently limited to virtual representation, simulation and offline monitoring, lacking closed-loop, real-time synchronization that supports proactive quality control (Christou et al., 2022; Onaji et al., 2022). Meanwhile, the integration of Large Language Models (LLMs) and API-driven interoperability remains fragmented: industrial systems rarely leverage natural language understanding, automated reasoning and structured data interpretation to support root-cause analysis, decision-making and knowledge dissemination across quality processes.

From a management viewpoint, SQM is evolving in a Quality 4.0 vision but still many of the solutions remain focused on TQM tools, Six Sigma tools or compliance driven tools and not fully on data-driven, intelligent and adaptative structures. Issues regarding integrating of operational data from the industry into a quality vision at the strategic level, making it difficult to sense variations at the appropriate time, make effective root cause analysis and take organizational learning from operational outcomes. Furthermore, the lack of standardized integration models between IoT sensing, digital twin dynamics and AI-driven reasoning creates siloed systems that restrict scalability, interoperability and consistent quality governance across smart factory environments.

Therefore, a clear need exists for unified, intelligent architectures that combine IoT data acquisition, dynamic digital twin synchronization, advanced analytics and LLM-based decision support within a Strategic Quality Management framework. Specifically, the **main open challenges** motivating this work can be summarized as follows:

1. Insufficient bidirectional real-time synchronization between physical production assets and their Digital Twin representations, limiting the granularity and responsiveness of continuous quality monitoring.
2. Architectural fragmentation across IoT, Digital Twin, and analytics layers, preventing the establishment of end-to-end, digitally integrated quality workflows within Smart Factory environments.
3. Underutilization of LLM and API capabilities as a semantic reasoning and integration layer for translating heterogeneous industrial data into actionable strategic quality intelligence.
4. Incomplete transition of SQM frameworks toward fully adaptive, learning-oriented Quality 4.0 architectures that support predictive and proactive quality governance at scale.

Addressing these challenges requires a holistic approach that transcends isolated technological or managerial solutions, creating a cohesive cyber-physical framework where data, simulation and intelligent reasoning jointly enhance strategic quality performance in modern

smart manufacturing.

1.3 Objectives and Research Questions

The overarching objective of this thesis is to identify, analyze, and formalize the research gap between established IoT-enabled Digital Twin systems for Strategic Quality Management in Smart Factories and the emerging integration of Large Language Model (LLM) and API architectures, and to propose and validate a conceptual framework for their integrated deployment.

1. Review the state of the art concerning IoT, Digital Twin, Smart Factories, and Strategic Quality Management in their current form, without LLM/API integration
2. Analyze the potential roles, value, and benefits that LLMs and APIs can bring to Strategic Quality Management in intelligent and data-driven manufacturing environments.
3. Identify and formalize the conceptual, technological, and practical gap between established IoT–Digital Twin–SQM systems and the integration of LLM/API technologies.
4. Explain the technical, organizational, and methodological barriers that hinder the development and industrial adoption of such integrated solutions.
5. Propose a conceptual framework for the integration of LLMs and APIs into IoT-Digital Twin-SQM systems, to address the identified gaps and optimize strategic quality management in Smart Factories.

To operationalize these aims and ensure a focused investigation, three research questions (RQs) have been developed, which directly align with the research objectives and address the core gaps identified:

RQ.1 *"How do IoT-enabled Digital Twin systems enable real-time data integration, process simulation and predictive analysis to support strategic quality decision-making in Smart Factories?"*

RQ.2 *"What strategic benefits arise from integrating LLM/API systems into IoT and Digital Twin-based SQM systems in Smart Factories in terms of efficiency?"*

RQ.3 *"What barriers limit the effective integration of LLM/API with existing IoT-Digital Twin SQM systems in Smart Factories, and how do they impact efficiency outcomes?"*

These questions will guide the systematic exploration of the research domain, ensuring that the study remains focused on addressing the identified gaps and achieving the stated research aims.

1.4 Literature Research Methodology

The literature underpinning this thesis was identified and selected through a Systematic Literature Review (SLR) protocol, a methodological framework designed to minimize selection bias and ensure transparent, reproducible identification of relevant scholarly contributions (A. Thomé et al., 2016). The protocol is structured into two sequential phases; each addressing a distinct dimension of the thesis's research domain and is described in full in the following subsections. (*Figure 1*)

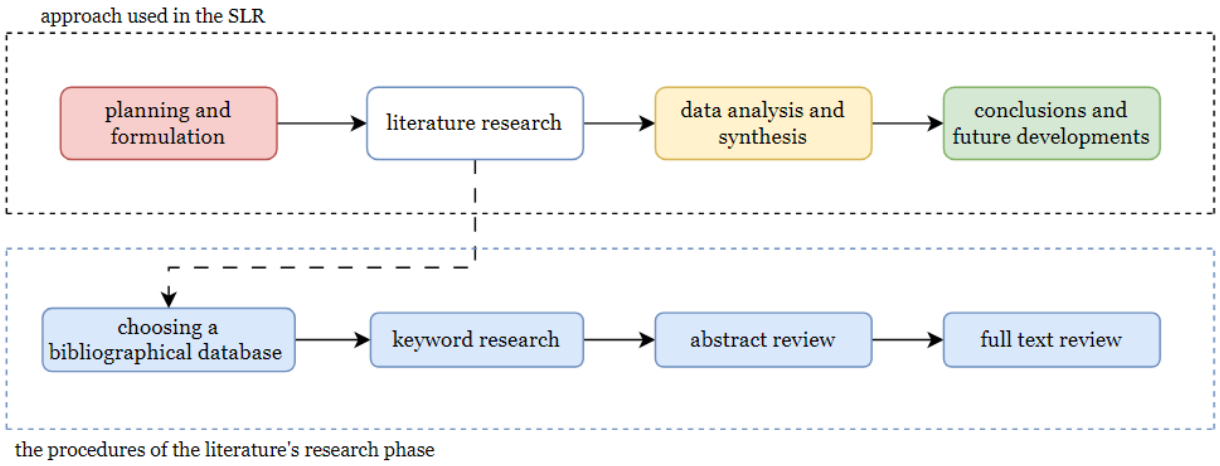

Figure 1 - Steps of the conducted study

The systematic review is designed to answer the three research questions formulated in *Section 1.3.* To do so, the following steps will be followed:

- Identification of relevant literature through database searching;
- Screening and selection of contributions based on predefined criteria;
- Quality assessment and data extraction;
- Synthesis of findings to highlight knowledge gaps;

Because of its thorough coverage and reliable indexing of peer-reviewed literature, Scopus, by Elsevier, was selected as the database for this thesis.

The search strategy was operationalized through a structured keyword query submitted to the Scopus database via its TITLE-ABS-KEY function. Keywords were organized into four semantic domains and combined using Boolean operators (AND, OR) to ensure both breadth and specificity. The resulting query was iteratively refined to balance recall (inclusion of all relevant studies) and precision (exclusion of irrelevant results), following established SLR best practice (A. Thomé et al., 2016).

Here are the discussed semantic areas:

→ *Digital Twin*
→ *The Internet of Things*
→ *Strategic Quality Management*
→ *Smart Factory*

The research scope was defined within the field of **industrial and manufacturing engineering**, focusing on digital transformation in smart factories. The topic was refined to emphasize:

- Industrial Internet of Things (IIoT)
- Digital Twin technologies
- Strategic Quality Management (SQM)
- Automation and process monitoring in smart factories
- Emerging AI tools (LLM/API) for advanced decision support

This scope ensures alignment with contemporary challenges in intelligent manufacturing and quality assurance.

Research Design

The systematic literature review is structured into **two sequential phases** to fully address the thesis focus:

IoT-enabled Digital Twin for Strategic Quality Management (SQM) in Smart Factories, and the research gap related to LLM/API architectures

1.4.1 Phase 1 – Targeted Review: IoT–DT–SQM

A targeted search was conducted to map existing literature on the integration of **IoT, Digital Twin, and Strategic Quality Management in Smart Factory environments**. The goal was to reconstruct the state of the art, technical applications, benefits, and limitations of these combined technologies for quality management.

The first search yielded **15 412** results, with the following query:

Code 1 – First Scopus Search String

"IOT" OR "INTERNET OF THINGS" OR "INDUSTRIAL IOT" AND "DIGITAL TWIN" OR "DIGITAL TWINS" AND "MANUFACTURING" OR "SMART MANUFACTURING" AND "QUALITY" OR "QUALITY MANAGEMENT" OR "QUALITY CONTROL" OR "STRATEGIC QUALITY MANAGEMENT"

Subsequently, Scopus's TITLE-ABS-KEY function was employed for the search; this function is specifically designed to retrieve only works where the selected keywords appear in the title, abstract, or keywords of each document in the database.

Approximate filters were then inserted into the databases used to obtain results relevant to the requirements of the paper; the filters applied are:

→ *Subject area*: engineering, business;
→ *Year of publication*: 2022 – *March* 2026;
→ *Language*: English

Following the application of these filters, the query was further refined through an iterative process to enhance its specificity and alignment with the study's core focus.

Query Refinement

To optimize the query formulation and ensure the inclusion of all relevant synonyms for the key concepts, *Doubao*, a generative AI-driven natural language processing tool, was employed. It was used to identify precise synonyms and related terms that align with the study's core themes, ensuring the query captured all relevant literature without introducing extraneous topics. Specifically, the LLM assisted in refining the terminology for the four key semantic domains: for IoT, it confirmed relevant synonyms such as "Industrial IoT (IIoT)" and "sensor networks"; for Digital Twin, it validated terms like "digital replica" and "virtual twin"; for SQM, it supported the inclusion of terms like "quality control" and "defect detection"; and for Smart Factory, it reinforced relevant phrasing such as "discrete manufacturing", "industrial production" and "production processes."

(It is acknowledged that the use of AI-assisted query refinement represents a non-standard methodological choice; all suggested synonyms were manually verified against the established literature on IoT-DT-SQM prior to inclusion in the final query string.)

The unfiltered search (limited to Engineering & Business, Articles/Reviews/Conferences) produced **230 results**. After applying strict filters, the final curated dataset resulted in **141 documents**, with the final query structured as follows:

Code 2 - Final Scopus Search String

```
TITLE-ABS-KEY ( ( "IOT" OR "INTERNET OF THINGS" OR "INDUSTRIAL IOT" OR "IIOT" OR "SENSOR NETWORKS" OR "SMART SENSORS" ) AND ( "DIGITAL TWIN" OR "DIGITAL TWINS" OR "VIRTUAL TWIN" OR "CYBER TWIN" OR "DIGITAL REPLICA" ) AND ( "MANUFACTURING" OR "SMART MANUFACTURING" OR "DISCRETE MANUFACTURING" OR "INDUSTRIAL PRODUCTION" OR "FACTORY AUTOMATION" OR "PRODUCTION PROCESSES" ) AND ( "QUALITY" OR "QUALITY MANAGEMENT" OR "QUALITY CONTROL" OR "STRATEGIC QUALITY MANAGEMENT" OR "SQM" OR "QUALITY ASSURANCE" OR "QUALITY IMPROVEMENT" OR "QUALITY PERFORMANCE" OR "DEFECT DETECTION" OR "QUALITY MONITORING" ) ) AND PUBYEAR > 2021 AND PUBYEAR < 2027 AND ( LIMIT-TO ( PUBSTAGE , "FINAL" ) ) AND ( LIMIT-TO ( LANGUAGE , "ENGLISH" ) OR LIMIT-TO ( LANGUAGE , "CHINESE" ) ) AND ( LIMIT-TO ( SUBJAREA , "ENGI" ) OR LIMIT-TO ( SUBJAREA , "BUSI" ) ) AND ( LIMIT-TO ( DOCTYPE , "RE" ) OR LIMIT-TO ( DOCTYPE , "AR" ) OR LIMIT-TO ( DOCTYPE , "CP" ))
```

Screening Process

The following passage is divided into two main steps:

→ *Abstract review*
→ *Full text review*

Selection standards applied to individual findings from the database's keyword research:

→ ***Inclusion***: Final published, full-text available documents; Subject area: Engineering, Business, Management, Publication year: 2022 to March 2026, English language (and Chinese for high-impact international works), Focus on IoT, Digital Twin, SQM, and Smart Manufacturing.

→ ***Exclusion***: documents about to be published, not accessible online, not in English, retracted documents, documents not belonging to the areas mentioned, published before 2022 and after March 2026.

Below is the *flowchart* for the procedure described above:

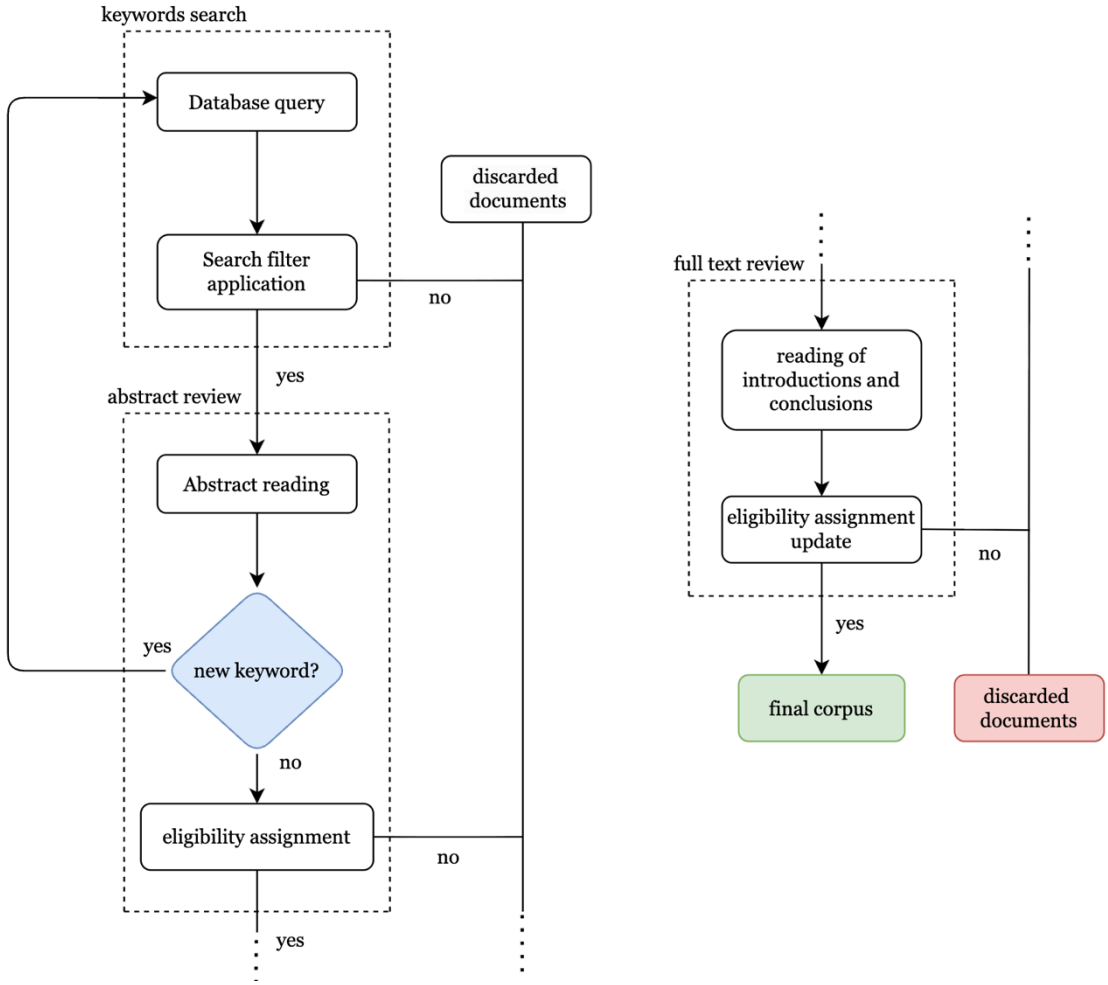

Figure 2 - Flowchart of the Systematic Literature Review

1.4.2 Phase 2 – Expanded Review: LLM/API in Broader Industrial Settings

Following the completion of Phase 1 (a focused state-of-the-art analysis on IoT, Digital Twin and Strategic Quality Management in Smart Factories), Phase 2 of the systematic literature review was launched as a **supplementary, two-step search strategy**. Initially, it aimed to directly investigate the integration of Large Language Models (LLMs) and API architectures within existing IoT–Digital Twin systems for Strategic Quality Management (SQM) in Smart Factories, core to formalizing the unique scientific gap of this thesis, which bridges foundational smart factory quality technologies and the emerging, understudied domain of advanced AI-driven interoperability and decision support.

To validate the initial research gap hypothesis, the refined Phase 1 search query (*Code 2)* was first expanded by adding LLM/API-related keywords and their established synonyms (including "large language models", "generative AI", "API integration", "application programming interfaces", "AI-driven decision support"). This targeted expansion yielded **zero valid results** on Scopus, confirming a complete absence of peer-reviewed literature explicitly linking LLM/API integration to IoT-Digital Twin-based SQM in smart factories, marking a definitive, unaddressed scientific gap.

To avoid misinterpreting this absence as a methodological flaw (rather than a genuine knowledge gap), the full, identical systematic literature review (SLR) protocol developed for Phase 1 was replicated in Phase 2, with one key adjustment: **the search scope was intentionally broadened and less restrictive** in keyword specificity, moving beyond the tightly targeted IoT–Digital Twin–SQM framework of Phase 1.

To comprehensively identify the research gap, **two research methods** were employed in Phase 2:

- The first method replicated the full systematic literature review (SLR) workflow as applied in Phase 1, with the key adjustment of a broadened search scope.
- The second method involved the application of Consensus AI, a scientific research platform designed to complement traditional database searches and enhance the comprehensiveness of the literature review.

For the **first research method** (the SLR workflow with the aforementioned Scopus query), all sequential steps of the SLR process (referenced back to the full protocol detailed in *Section 1.4*) were repeated to ensure methodological rigor and avoid false gap identification due to overly rigid search parameters. These steps included: database selection (retaining Scopus for consistency), keyword synonym expansion, iterative query refinement, application of consistent publication filters (2022 – March 2026, English/Chinese, engineering/business subject areas), abstract screening, full-text review, and quality assessment.

Code 3 – Research Gap Scopus Search String

```
TITLE-ABS-KEY (("LARGE LANGUAGE MODEL" OR LLM OR "GENERATIVE AI" OR
"FOUNDATION MODEL") AND ("API" OR "API INTEGRATION" OR "REST API") AND
("MANUFACTURING" OR "SMART FACTORY" OR "SMART MANUFACTURING" OR
"INDUSTRY 4.0"))
```

The **second research method** used was using Consensus AI, a science research tool that uses artificial intelligence to sum up the findings of peer-reviewed papers. This replicated process was critical to confirm that the lack of relevant literature was not a result of flawed search execution, but rather a genuine absence of scholarly work on this specific integration.

The core objectives of this replicated, broadened search were fourfold:

- Verify whether any conceptual, technical or empirical frameworks exist that combine LLM/API interoperability with IoT-Digital Twin architectures for strategic quality governance;
- Identify technical, operational and managerial limitations documented in broader AI-manufacturing literature that may apply to this untested integration;
- Highlight missing practical applications, performance metrics and implementation guidelines for AI-enhanced smart factory quality systems;
- Formalise the unique, original scientific gap that forms the foundational rationale for this thesis, distinguishing it from existing literature on standalone Digital Twin, IoT or generic AI manufacturing applications.

1.5 Summary of Identified Research Gaps

The two-phase systematic literature review described in *Sections 1.4.1* and *1.4.2* reveals a consolidated set of research gaps that currently hinder the effective deployment of IoT-enabled Digital Twin systems for Strategic Quality Management (SQM) in Smart Factories. While the individual technologies, IoT, Digital Twins, and advanced analytics, have reached considerable maturity, their integration into a coherent, human-centric quality governance framework remains incomplete. Five primary gaps emerge from the analysis, which are examined in detail in *Chapter 2*, *Table 3*.

First, current DT-IoT systems have interface limitations as quality data is being represented as complex dashboard so that only expert will be able to understand and there is not available interface to provide this data into human-like language. Secondly heterogeneous communication protocols, outdated PLCs cause the fragmentation of cyber physical orchestrations. Third, *in* current DT platforms, there is no available reasoning layer that allows for multi-step causal analysis to identify the root cause or to generate preventive actions. Fourth, knowledge accessibility is limited because quality engineers cannot interactively query DT models without specialized software skills. Fifth, there is a standardization deficit: no unified architectural framework exists for integrating Large

Language Models (LLMs) with IoT-Digital Twin ecosystems specifically for SQM.

These five gaps collectively motivate the conceptual framework proposed in *Chapter 3*. The framework is explicitly designed to address each gap by introducing a Generative AI Layer (LLM-based semantic reasoning), a unified API orchestration architecture, and a closed-loop human-supervised governance structure tailored to the requirements of SQM in Smart Factory environments.

1.6 Thesis Structure

Chapter 2 presents a comprehensive review of the state of the art, structured thematically from the foundational convergence of cyber-physical systems and quality management, through the current landscape of IoT-DT industrial applications, to the emerging research gaps identified at the intersection of interpretability, cyber-physical orchestration, and LLM-augmented quality governance. The chapter concludes by formalizing the five research gaps that directly motivate the proposed framework.

Chapter 3 introduces the LLM-DT-SQM framework, a five-layer closed-loop architecture integrating IoT-enabled Digital Twins with Large Language Model capabilities for Strategic Quality Management in Smart Factories. This chapter defines a complete specification of the conceptual architecture of the framework, the justification behind the layer-by-layer design, the functional modules, the assembly of the context window, the governing system and the methodical cycle of the framework which represents the actual research proposed by this thesis.

Chapter 4 presents the industrial case study, a precision CNC machining line for automotive powertrain cylinder heads, through which the framework is instantiated layer by layer and validated through a simulation-based reference implementation. The case study demonstrates the framework's operational logic across representative quality scenarios, including real-time anomaly detection, automated root-cause analysis, prescriptive CAPA generation, and closed-loop feedback, supported by a functional proof-of-concept prototype.

Chapter 5 presents the experimental results and evaluation, assessing framework performance across four test scenarios using twelve quantitative KPIs distributed across quality intelligence accuracy, system performance, operational impact, and governance compliance. Results are compared against conventional SQM practice to quantify the framework's incremental value and identify its primary limitations.

Chapter 6 draws conclusions, answers the three research questions with reference to the empirical results, acknowledges the structural limitations of the simulation-based evaluation, and identifies future research directions including multi-agent LLM orchestration, on-premises deployment, and cross-factory validation studies.

2 Literature Review

2.1 Introduction: The Convergence of Cyber-Physical Systems and SQM

The ongoing digital transformation of manufacturing, driven by the Industrial Internet of Things (IIoT) and Digital Twins (DTs), is driving a paradigm shift that is reshaping self-governed Cyber-Physical Production Systems (CPPS). These sophisticated systems support adaptive and dynamic configurations for collaborative operations, thereby reconfiguring the paradigms through which quality is managed and optimized across production ecosystems (Redchuk & Walas Mateo, 2025). The most important feature of this shift is the adoption of Cyber-Physical Systems (CPS) into manufacturing environments, which represents a fundamental departure from traditional, siloed quality management approaches. In this new context, Strategic Quality Management (SQM) is no longer confined to localized inspection tasks; instead, it emerges as a distributed intelligence function that orchestrates quality objectives across the entire production network.

By mirroring physical assets into a digital environment, factories can synchronize production variables with high-level quality objectives, ensuring that every operational shift is aligned with the strategic goal of zero-defect manufacturing. SQM evolves into a cross-cutting coordination and decision layer that exploits CPS-generated data and digital twins to align operational control with long-term quality and performance targets. Digital twin–enabled environments gather heterogeneous process and quality data into centralized or cloud-based platforms and use data-driven and physics-based models to "optimize operations" while balancing energy efficiency, operational costs, and product quality in closed-loop control architectures (Onaji et al., 2022). As a result, SQM becomes an embedded, system-level function that leverages CPS connectivity and digital representations to continuously coordinate local decisions with strategic zero-defect objectives across the manufacturing network.

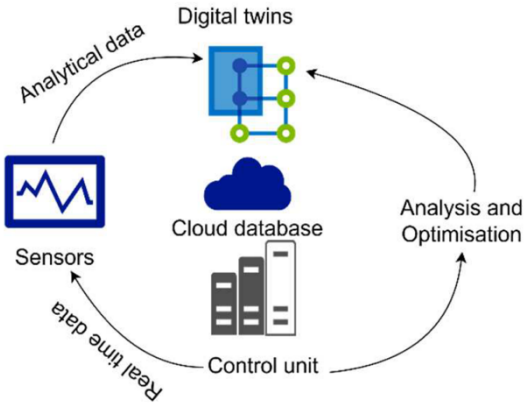

Figure 3 - Interaction of real-time data using cloud database and digital twins

2.1.1 Technical Foundations of IoT-Enabled Digital Twins

This section focuses on the specific technical requirements for IoT-enabled DT in SQM contexts. Modern industrial architectures increasingly adopt a distributed, multi-layer Edge–Fog–Cloud computing pipeline to enable robust, real-time IoT-enabled Digital Twin (DT) operations. In this layered framework, physical shop-floor assets are equipped with industrial sensors and actuators that continuously capture critical operational variables, including vibration, pressure, temperature, speed, position, and energy consumption. These real-time data streams are transmitted securely via IIoT communication protocols to cloud-based storage and processing platforms, where they are cleaned, aggregated, and fused to feed high-fidelity digital twin models for dynamic simulation, predictive analysis, and closed-loop process control (Tiwari et al., 2025).

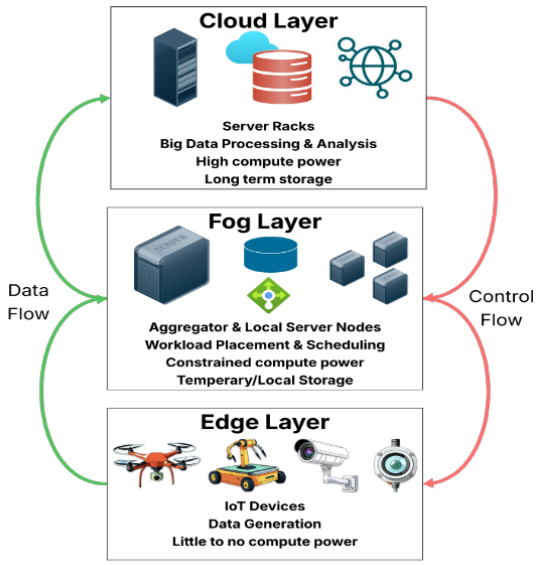

Figure 4 - High-level of edge, fog, and cloud layers

To overcome latency, bandwidth, and privacy limitations associated with exclusive cloud reliance, edge and fog computing layers are strategically integrated to execute low-latency analytics, local decision-making, and time-critical anomaly detection directly near the data source. This hierarchical structure ensures that mission-critical quality monitoring and fault identification tasks comply with strict Quality of Service (QoS) constraints, while less time-sensitive computations and long-term data storage are offloaded to the cloud. Within this architecture, digital twin mechanisms simulate workload distribution, resource contention, and computational utilization across heterogeneous fog nodes, enabling accurate forecasting of response times and supporting reliable, QoS-aware workload placement (Swaminathan et al., 2025). By combining IIoT connectivity, distributed computing, and synchronized virtual replicas, IoT-enabled digital twins establish a continuous bidirectional data flow between physical systems and their virtual counterparts. This synchronization enables real-time state updates, autonomous model adaptation, and proactive simulation of process deviations,

forming the technological foundation for responsive, predictive, and autonomous quality management in smart factories (Bongomin et al., 2025).

As digital twin paradigms evolve from isolated asset models to factory-wide ecosystems, they increasingly consolidate diverse upstream and downstream data into unified "data continents" to support collaborative optimization of process parameters and quality-critical variables across entire production lines. This integrated structure breaks down long-standing information silos between consecutive manufacturing stages, avoiding the limitations of traditional single-process models that neglect inherited process influences (J. Li et al., 2024). Supported by IIoT infrastructures, heterogeneous data from physical assets, controllers, and inspection systems are aggregated into centralized data layers, forming a reliable foundation for real-time simulation and model training. The Edge–Fog–Cloud architecture thus constitutes the enabling technical infrastructure for the predictive quality governance capabilities examined in the following section.

2.2 Predictive Maintenance as Strategic Quality Tool

Within the framework of Strategic Quality Management (SQM), Predictive Maintenance (PdM) emerges as a structured closed-loop quality governance mechanism rather than a marginal, secondary support function limited to equipment reliability. It supports measurable improvements in industrial performance metrics, including reduced unscheduled downtime, enhanced Overall Equipment Effectiveness (OEE), and lower rates of non-conforming products (Christou et al., 2022). From this strategic perspective, the value of PdM extends far beyond simply preserving machine availability or extending asset lifecycles; it functions as a proactive quality safeguard that continuously monitors process stability, identifies subtle deviations from nominal operating conditions, and enables targeted corrections before process drift escalates into irreversible quality losses. In this sense, predictive maintenance becomes an integral part of quality assurance, embedding reliability considerations directly into the governance of product conformance.

This paradigm shift is further reinforced by recent advancements in digital twin (DT) research, which consistently emphasize that the accuracy, real-time synchronization, and structural consistency of digital twin models directly determine the efficiency of manufacturing execution, the stability of product quality, and the overall reliability of automated production systems (Sun et al., 2025). As underlined in studies focused on smart assembly automation, digital twins operate as dynamic replicas driven by continuous streams of real-time operational data, enabling full-cycle monitoring, in-depth performance analysis, and timely intervention across every phase of production activity (Webb et al., 2025).

Collectively, these contributions support a fundamental conceptual reorientation: early warning signals originating from production equipment, including abnormal vibration patterns, progressive tool wear, positioning inaccuracies, sensor drift, and pressure or temperature fluctuations, must be systematically classified as quality-related risks rather than isolated technical disturbances relevant only to maintenance departments. Through this modification

to interpretation, quality concerns now also drive operation at the earliest perceivable indications of process drift. Empirical evidence that this consolidated logic applies to shop-floor operations is abundant in real-world industrial examples. Adaptive digital twin implementations in precision milling and drilling processes monitor tool wear, thermal expansion, process anomalies, while achieving, for instance, higher precision in holes, greater part stability, and better end-part quality (Savazzi et al., 2025). Similarly, dedicated tool wear monitoring systems are intentionally designed to minimize scrap generation, preserve surface integrity, and guarantee compliance with strict dimensional specifications for finished components (Hu et al., 2022). Parallel advancements in quality inspection architectures employ semantic digital twin models that support flexible parameter adjustment, real-time configuration updates, and adaptive control logic, ensuring consistent and repeatable quality delivery even within high-mix, flexible production environments characterized by frequent product changes (Hu et al., 2022). In each industrial application, maintenance-derived intelligence is systematically translated into actionable quality intelligence, effectively preventing dimensional drift, surface defects, structural non-conformities, and variability in material properties at the earliest processing stage.

As noted in hybrid production management research, digital twins supply the continuous real-time data, high-fidelity simulation capabilities, and predictive visibility required for informed, optimal decision-making, while artificial intelligence algorithms support automated yet human-validated corrective actions that balance speed, precision, and operational safety (Schmid & Winkler, 2022). Within this integrated formulation, predictive maintenance transcends narrow concerns of operational efficiency and cost reduction to become an indispensable structural component of the enterprise-wide quality architecture. By synchronizing equipment health monitoring, process repeatability enforcement, and final product conformity assurance within a unified, closed-loop control system, PdM elevates quality from a downstream inspection activity to a foundational, end-to-end governance principle embedded within every layer of production operations.

Predictive quality models embedded within DT architectures leverage a variety of machine learning paradigms. Supervised learning approaches, such as gradient boosting and deep neural networks, are trained on historical quality records to predict defect probabilities and non-conformance likelihood before they manifest on the production line (Kabir et al., 2025).

Nevertheless, the current generation of predictive DT-SQM systems faces a fundamental bottleneck: predictive outputs are typically presented to quality engineers through numerical dashboards or static alert systems that require specialized domain knowledge to interpret and act upon. This interpretability gap between sophisticated predictive models and actionable strategic decisions represents a critical limitation that motivates the integration of higher-order AI capabilities, specifically, Large Language Models, as discussed in the following sections of this chapter.

2.2.1 Big Data Analytics

The transformation of raw manufacturing data into actionable quality intelligence constitutes the foundational value proposition of Big Data Analytics (BDA) within IoT-enabled SQM systems. The volume, velocity, and variety of data generated by modern Smart Factories, spanning sensor telemetry, machine logs, vision inspection outputs, and supply chain records, far exceed the processing capacity of traditional statistical process control (SPC) methods (Christou et al., 2022; Kabir et al., 2025). BDA frameworks address this challenge by providing scalable data ingestion, processing, and analytical infrastructure that operates in close coupling with DT platforms and IoT networks.

A layered architecture for BDA-enabled quality intelligence, as synthesized from the reviewed literature, comprises five functional strata: (i) data ingestion via IIoT sensors and industrial communication protocols (OPC-UA, MQTT); (ii) persistent storage in time-series databases and data lakes; (iii) stream and batch processing for SPC computation and anomaly detection; (iv) AI/ML-based predictive analytics; and (v) decision-support integration with DT dashboards, MES, and ERP systems.

Table 1 summarizes these layers and their associated quality intelligence outputs.

Table 1 - Layered Big Data Analytics Architecture for IoT-DT Quality Intelligence

Layer	Key Technologies	Quality Intelligence Output
Ingestion	IIoT Sensors, OPC-UA, MQTT	Real-time process parameters (temperature, vibration, torque)
Storage	Data lakes, time-series DBs, cloud/edge	Historical quality records, traceability logs
Processing	Apache Spark, stream processing, ML pipelines	SPC charts, anomaly scores, process fingerprints
Analytics & AI	DNN, LSTM, Gradient Boosting, DRL	Predictive defect rates, root-cause scores, yield forecasts
Decision Support	DT dashboards, KPI engines, ERP integration	Corrective action recommendations, strategic quality KPIs

From a strategic management perspective, BDA within DT-IoT systems contributes to quality intelligence in three analytically distinct ways: descriptive analytics (what happened, historical process reconstruction and defect root-cause tracing); predictive analytics (what will happen, failure forecasting and yield prediction); and prescriptive analytics (what should be done, optimal process parameter recommendations and maintenance scheduling) (Bongomin et al., 2025; Tiwari et al., 2025). The integration of all three analytical modes within a unified

DT platform represents the current frontier of SQM, yet the translation of these analytics into actionable, interpretable decisions remains a significant research and practical challenge.

2.3 Current Industrial Applications of DT-IoT for SQM

The integration of Digital Twins and IoT networks for Strategic Quality Management has progressed substantially from theoretical frameworks to validated industrial implementations across a wide range of manufacturing sectors. In this section, the current status of deployment based on the empirical data obtained through the systematic literature review is analyzed and presented. Both, the identified proved capabilities and unaddressed issues within operational DT-IoT SQM systems are highlighted.

Validated deployments of IIoT-DT integrations confirm that tangible SQM benefits result across different architectures in the aerospace, shipyard and SME manufacturing industries. Two of the aerospace CNC applications resulted in validated IIoT-DT parameter conformance system configurations with real-time cross-reference of PDM/ERP reference data conforming to AS9100 standard. At enterprise scale, the Fincantieri Digital Twin program instantiated a DT architecture spanning eight thematic domains including an explicit Quality domain. For SME manufacturers, BOM-integrated DT frameworks have enabled PLM-aligned quality work instructions through supplier-collaborative process design. Key architectures, implementation contexts, and quality outcomes are summarized in *Table 2.*

Table 2 - Cross-Sector Industrial Applications of DT-IoT for SQM

Reference	Sector	DT-IoT Application	SQM Outcome
Christou et al. (2022)	Manufacturing	End-to-end IIoT platform for Quality 4.0	Improved OEE; zero-defect production
Sokolov et al. (2023)	Aerospace CNC	IIoT-DT parameter compliance control (AS9100)	Real-time conformance; deviation control
Khdoudi et al. (2024)	Plastic Injection	DRL-based DT for autonomous process optimization	Reduced defects; minimal human intervention
Fattahi et al. (2025)	Grinding	Expert system adaptive DoE with NSGA-II	>90% validation accuracy; surface defect prevention
(Pilutti Namer & Tornielli, 2025)	Shipbuilding	Enterprise DT (8 domains incl. quality, H&S)	End-to-end supply chain quality visibility
Tonelli et al. (2026)	SME Manufacturing	BOM-integrated 3D DT with PLM-aligned QM	Collaborative quality work instructions
Tiwari et al. (2025)	Manufacturing	Cloud DT vs. traditional process monitoring	Quantified failure rate and downtime reduction

Across these implementations, several common architectural patterns emerge. First, IoT data collection is typically mediated by industrial communication protocols (OPC-UA, MQTT, AMQP) that feed time-series data to cloud or edge-based DT computation layers. Second, ML-based predictive models work within the DT structure to produce quality forecasts and anomaly alerts. Third, connectivity to enterprise systems (ERP, MES, PLM) will create the bi-directional coupling of quality information to operation and strategic management processes. Fourth, visualizations and reports (mainly dashboards) convey the quality intelligence information to engineering and management levels.

However, although these achievements, a main issue identified within the literature is lack of an intelligent, interactive interface layer that would translate the multi-dimensional and multiple outputs generated by a DT-IoT SQM system into understandable quality management decisions to non-experts. As (Galdino et al., 2025) observe, existing DT platforms present quality data that is structurally rich but cognitively demanding, requiring specialized expertise to extract strategic insights. This gap, the interface and interpretability bottleneck, is addressed in the following section.

2.4 Emerging Challenges and Research Gaps

Despite the considerable advances documented in the preceding sections, the current generation of IoT-enabled Digital Twin systems for SQM faces a set of structural and technical challenges that constrain their strategic effectiveness. These challenges are not merely engineering problems to be resolved through incremental improvement; they represent fundamental gaps between the current capability frontier and the strategic quality management requirements of modern Smart Factories. This section identifies and analyses the two primary challenge clusters: the interface and interpretability bottleneck, and the fragmentation of cyber-physical orchestration.

2.4.1 The Interface and Interpretability Bottleneck

The first and arguably most consequential challenge concerns the interface between the analytical power of DT-IoT SQM systems and the human actors responsible for quality governance decisions. Contemporary DT platforms are architecturally optimized for data richness and computational depth, yet typically lack the interface capabilities to communicate insights to stakeholders who lack specialized technical backgrounds (Galdino et al., 2025; Ouerghemmi & Ertz, 2025).

This challenge manifests in several interconnected forms. From a cognitive load perspective, quality managers confronted with multi-dimensional DT dashboards, displaying simultaneous time-series for hundreds of process parameters, statistical process control charts, predictive alerts, and simulation outputs, face an information overload that paradoxically undermines the strategic decision-making value of the system (Xia et al., 2025). The cognitive architecture of current DT interfaces presupposes a highly trained expert user capable of interpreting complex visualizations, cross-referencing data streams, and translating analytical outputs into quality

management actions, a profile that is increasingly misaligned with the broader operational reality of Smart Factory quality teams.

From an interpretability perspective, the ML and DRL models embedded within DT-SQM systems operate as largely opaque functions: they generate predictions and recommendations with high accuracy but without providing human-comprehensible explanations of the causal reasoning that underpins those outputs (Bieniek et al., 2025; Wang et al., 2024). This opacity is not merely an intellectual inconvenience; it constitutes a practical barrier to trust-based adoption and a regulatory compliance risk in contexts governed by standards such as ISO 9001:2015, which requires documented evidence of quality reasoning and decision traceability (Galdino et al., 2025).

Furthermore, the inability to query DT systems in natural language, the modality in which quality professionals naturally formulate questions ('What caused the increase in surface defect rate during the night shift on line 3?'), forces quality analysts to reformulate their investigative questions into the constrained query languages of specific platforms. This translation overhead constitutes a substantial organizational friction that reduces the operational value of DT investments and contributes to the underutilization of available quality intelligence (Freire et al., 2024; Lou et al., 2025).

2.4.2 Fragmentation in Cyber-Physical Orchestration

The second major challenge cluster concerns the technical fragmentation that characterizes the cyber-physical architecture of most current DT-IoT SQM deployments. The consequence of fragmentation challenges for SQM is significant: the DT cannot maintain a coherent, unified quality picture of the production system when its data inputs are inconsistently structured, unreliably timed, or semantically incompatible. This undermines the predictive accuracy of quality models, the reliability of automated alerts, and the fidelity of simulation-based scenario analysis (Ghofrani et al., 2025).

Legacy system compatibility represents a specific and particularly intractable dimension of fragmentation. The majority of industrial IoT deployments must operate alongside legacy PLCs and automation equipment that were designed before the era of cloud connectivity, open APIs, and standardized data exchange (Doğruluk & Acikgoz, 2025). Retrofitting these systems for DT integration without disrupting production continuity is a substantial engineering and organizational challenge that is only partially addressed by current middleware solutions.

These fragmentation challenges collectively prevent the development of the integrated, semantically coherent causal representation that advanced DT-SQM reasoning requires, the structural gap that motivates the architectural integration proposed in *Chapter 3*.

Table 3 presents a consolidated view of the key research gaps identified across the reviewed literature.

Table 3 - Research Gaps in DT-IoT SQM Systems

Research Gap	Description	Relevant Literature
Interface & Interpretability	DT dashboards present raw data that requires expert interpretation; no natural language layer for non-specialist users	Galdino et al. (2025); Ouerghemmi & Ertz (2025)
CPS Fragmentation	Heterogeneous protocols (OPC-UA, MQTT, REST) and legacy PLCs hinder seamless data orchestration across subsystems	Doğruluk & Acikgoz (2025); Ghofrani et al. (2025)
Absent Reasoning Layer	DT platforms lack multi-step causal reasoning needed for root-cause analysis and corrective action generation	Xia et al. (2025); Wang et al. (2024)
Knowledge Accessibility	Quality engineers cannot interactively query complex DT models without specialized software skills	Freire et al. (2024); Lou et al. (2025)
Standardization Deficit	No unified architectural framework for integrating LLMs, DTs, and IoT in SQM contexts	Galdino et al. (2025)

This analytical diagnosis motivates the core research contribution of this thesis: the proposition that Large Language Models, deployed via structured API architectures, can serve as the semantic integration and reasoning layer that bridges the gap between the data richness of current DT-IoT SQM systems and the knowledge-driven quality intelligence that Strategic Quality Management demands. This conceptual bridge is developed in the sections that follow.

2.5 Bridging the Gap: LLM-Enhanced DTs for SQM in Smart Factories

The preceding analysis identifies a structural gap: current DT-IoT SQM systems are data-rich and computationally powerful, yet cognitively opaque and dependent on specialist expertise to interpret. This section develops the case that Large Language Models, integrated via structured API architectures, constitute the natural interface and reasoning layer to bridge this gap, transforming data-rich DT platforms into knowledge-accessible quality governance.

Although classical DT-IoT SQM approaches typically deliver highly rich information, deep and intensive computations, strong analytic capacities, their deficiencies include low interpretability, poor usability by interface based on complicated procedures, limited multi-step reasoning ability, especially the key to decision making based on qualitative strategy. The modern generative LLM with their extensive knowledge, formed over large corpora of technical-scientific and operational text, complement the current lack of interpretability-simplicity and ability to perform multi-step causal inference required to aid in making strategy-quality decisions effectively and coherently-in an understandable manner (Lou et al., 2025; Wang et

al., 2024; Xia et al., 2025).

This complementarity defines the architectural logic of LLM-enhanced DTs for SQM. The DT provides the real-time, physics-grounded, data-synchronized digital representation of the production system; the LLM provides the semantic reasoning and natural language interface layer that transforms this representation into actionable quality management intelligence. The API architecture, discussed in *Section 2.5.4*, provides the integration backbone that enables bidirectional communication between these two components and between the integrated system and the broader ecosystem of IoT devices, enterprise systems, and knowledge repositories.

Evidence from the emerging literature on LLM applications in manufacturing and cyber-physical systems supports this integration thesis. (Gill et al., 2025) demonstrated that LLM agents connected to DT simulations of a process plant could autonomously generate fault diagnosis reports and corrective action plans with minimal human prompting, achieving effective fault mitigation that would previously have required hours of expert analysis. (Lou et al., 2025) proposed a five-dimensional 'cognitive DT' framework for Industry 5.0 in which a LLM agent provides the cognitive intelligence layer, responsible for contextual understanding, task planning, and adaptive incident response, while the DT layer provides real-time environmental state. (Warburton, 2025) described a vision of 'hyper adaptive' production systems in which cognitive DTs powered by LLMs enable autonomous production scheduling and quality decision-making in response to dynamic shop-floor conditions.

The SQM-specific implications of this integration are profound. Quality audits that currently require specialist engineers to manually interrogate multiple data systems and reconcile complex analytical outputs become conversational interactions in which quality managers pose questions in natural language and receive comprehensive, evidence-grounded analytical responses. Root-cause analyses that currently take days to conduct become automated inference processes in which the LLM-enhanced DT traces the causal chain from quality deviation to contributing process factors, cross-references relevant quality standards and historical precedents, and generates a structured corrective action report. These capabilities are examined in detail in *Sections 2.6.1* and *2.6.2.*

Current DT-SQM dashboards impose high extraneous cognitive load through several mechanisms: the simultaneous display of large numbers of process parameters without automatic relevance ranking; the use of statistical representations (control charts, histograms, scatter plots) that require statistical literacy to interpret correctly; the absence of contextual explanations that relate current process states to quality implications; and the need to navigate multiple screens, modules, and data views to assemble a complete picture of a quality event or trend (Galdino et al., 2025; Ouerghemmi & Ertz, 2025). These demands are particularly pronounced in complex, multi-line manufacturing environments where quality events may involve the interaction of dozens of contributing parameters across multiple process stages.

LLM-powered conversational interfaces address these cognitive load challenges by serving as intelligent intermediaries that perform the cognitive work of data navigation, relevance filtering, and representational translation on behalf of the user. When a quality manager poses

the question 'Why did the surface roughness on Line 4 exceed the upper specification limit between 02:00 and 05:00 this morning?', an LLM connected to the DT-IoT data layer can autonomously retrieve the relevant sensor data for the specified time window, identify the parameters that deviated from their normal ranges, cross-reference these deviations with known process-quality correlations, and synthesize the findings into a natural language explanatory narrative, dramatically reducing the cognitive work required of the user (Gill et al., 2025; Meng et al., 2025).

Empirical support for the cognitive load reduction benefits of LLM-based quality interfaces is provided by (Freire et al., 2024), who demonstrated in a user study that LLM-powered knowledge management tools in a manufacturing context enabled faster information retrieval and improved issue resolution compared to traditional search-based interfaces, even when users expressed initial preferences for human expert consultation. (Chowdhury, 2025) reported a 30% reduction in MES configuration and setup time through LLM-driven recommendation systems, suggesting that the cognitive efficiency gains generalize beyond quality-specific applications to broader manufacturing intelligence tasks.

2.5.1 Large Language Models (LLMs) as Orchestrators of Industrial Knowledge

Industrial quality management is an inherently knowledge-intensive activity that draws on multiple information categories simultaneously: real-time process sensor data, historical quality records, engineering specifications and tolerances, material properties and supplier certifications, quality standards and regulatory requirements, maintenance history and equipment condition data, and tacit expert knowledge encoded in procedures, work instructions, and institutional memory (Galdino et al., 2025; Wang et al., 2024). Existing DT-SQM platforms can manage and correlate the first two knowledge types, real-time operational data and historical operational data, but fall short in their ability to accommodate and reason with additional types of knowledge, residing in textual documents, internal databases and human expertise.

It is LLMs that can potentially fill this gap, especially if supported by Retrieval-Augmented Generation (RAG) architectures and knowledge graphs. RAG enables LLMs to retrieve and incorporate relevant information from structured and unstructured knowledge repositories, including quality standards documents, engineering manuals, and historical root-cause analysis reports, at inference time, grounding their analytical outputs in verified domain knowledge rather than relying solely on parametric knowledge encoded during pre-training (Lou et al., 2025; Meng et al., 2025). Knowledge graph integration provides the explicit relational structure, encoding relationships between process parameters, equipment components, quality characteristics, and failure modes, that enables LLMs to perform structured causal reasoning rather than merely pattern-matching over text (Meng et al., 2025; Videsjorden et al., 2025).

The orchestration role of LLMs in this context extends to multi-agent architectures, in

which specialized LLM agents are assigned responsibility for different knowledge domains or analytical functions, process monitoring, root-cause analysis, compliance checking, report generation, and coordinate their outputs through a central orchestrator agent that synthesizes the results into a unified quality management response (Lim et al., 2024; Xia et al., 2025). This multi-agent paradigm mirrors the organizational structure of quality management teams and enables the LLM layer to scale its capabilities to the complexity of the quality management task at hand.

The transformative potential of LLMs for DT-SQM is fundamentally contingent on their ability to interact dynamically with external data sources, computational tools, and enterprise systems. An LLM operating in isolation, without access to real-time sensor data, DT simulation outputs, ERP quality records, or knowledge repositories, is limited to the static parametric knowledge acquired during pre-training, which is insufficient for the time-sensitive, context-specific demands of industrial quality management. Application Programming Interface (API) architectures constitute the technical infrastructure that transforms LLMs from static knowledge systems into dynamic, action-capable agents embedded in the living fabric of Smart Factory operations (Luo et al., 2025; Xia et al., 2025).

2.5.2 From Static Knowledge to Real-Time Action: The Role of Function Calling

The mechanism that operationalizes API-based LLM action in IoT-DT SQM systems is function calling, a capability, available in contemporary LLM APIs (including OpenAI GPT-4, Anthropic Claude, and Google Gemini), that enables the LLM to invoke externally defined functions with structured parameter specifications in response to user queries or automated analytical triggers (Cui et al., 2023; Xia et al., 2025). Function calling represents the critical technical bridge between the LLM's natural language reasoning capabilities and the deterministic, data-precise operations required for industrial quality management.

In the DT-SQM context, function calling enables the following operational sequence: (i) the quality professional or automated monitoring system poses a query or triggers an analysis event; (ii) the LLM interprets the semantic content of the query and identifies the data sources, computations, and actions required to address it; (iii) the LLM generates structured function call specifications, including function names and parameter values, for the required API operations; (iv) the API layer executes the specified functions against the connected DT-IoT infrastructure; (v) the results are returned to the LLM; and (vi) the LLM synthesizes the results into a natural language response or automated action (Cui et al., 2023; Wang et al., 2024). This sequence transforms the LLM from a text-generation system into an intelligent orchestrator of industrial data and quality management workflows.

Table 4 presents a classification of API function types and their associated SQM applications in the DT-IoT context.

Table 4 - API Function Types and their SQM Applications

API Function Type	Industrial Role	Example Action
Sensor query functions	Pull live IoT readings from DT data broker	get_sensor_data(sensor_id, time_range)
Quality alert triggers	Invoke SPC alarm or NCR workflow via API	trigger_quality_alert(param, threshold)
Simulation calls	Run DT 'what-if' scenarios from NL prompt	run_dt_simulation(scenario_config)
ERP/MES integration	Create non-conformance reports or work orders	create_work_order(process_id, action)
Knowledge retrieval (RAG)	Retrieve ISO/AS standard clauses for context	retrieve_knowledge(query, standard='ISO9001')

The transformative implication of function calling for SQM is that quality intelligence becomes action-coupled: an LLM that identifies an emerging quality deviation does not merely report it as text but can autonomously trigger the appropriate quality management workflow, generating an NCR, scheduling a maintenance intervention, adjusting a process parameter within predefined safe bounds, or alerting the responsible quality engineer, through structured API calls (Gill et al., 2025; J. Li et al., 2025). This coupling of reasoning and action is the essential enabler of autonomous quality management.

2.5.3 API-Centric Design: Connecting the "Brain" (LLM) to the "Senses" (IoT/DT)

An API-centric design philosophy for LLM-enhanced DT-SQM systems prioritizes the architectural principle that every data source, analytical function, and action capability within the Smart Factory should be exposed as a well-documented, standardized API endpoint accessible to the LLM orchestration layer. This principle, analogous to the service-oriented architecture (SOA) paradigm in enterprise software engineering, ensures that the LLM can access any required capability without requiring custom integration code for each specific interaction (Manwatkar, 2025; Xia et al., 2025).

In practice, API-centric design for LLM-DT-SQM systems involves three categories of API exposure. IoT and sensor APIs provide real-time and historical access to process parameter data, quality measurement results, machine condition telemetry, and environmental monitoring readings from the factory floor, typically mediated by IoT data brokers and DT data services. DT simulation and analytics APIs expose the computational capabilities of the DT layer, including physics-based simulation, ML model inference, SPC calculation, and digital scenario

analysis, as callable functions that the LLM can invoke to generate analytical outputs grounded in the DT's real-time process representation. Enterprise system APIs provide integration with MES, ERP, PLM, and quality management systems (QMS), enabling the LLM to retrieve quality records, access engineering specifications, create non-conformance reports, and trigger corrective action workflows in response to analytical findings (Manwatkar, 2025).

2.5.4 Architectural Synergy: Integrating LLMs into the DT Framework

The integration of LLMs into DT frameworks for SQM encompasses several architectural patterns, each reflecting different balances between cognitive centralization, real-time responsiveness, and organizational scalability. Drawing on the emerging architectural literature, particularly the systematic frameworks proposed by (Lou et al., 2025; Luo et al., 2025; Warburton, 2025; Xia et al., 2025), this section identifies and characterizes three primary integration architectures.

The simplest and most easy to manage integration architecture is LLM-as-Interface; here the LLM is used as a conversational interface to an existing DT-SQM platform. So the LLM essentially sits on top of the current DT-SQM infrastructure but uses the full range of data and analysis capabilities of that DT. In this scenario natural language requests are translated into an API call on the DT level, data retrieved from this and translated into a natural language response.This architecture is well-suited to near-term deployment scenarios, as it does not require fundamental changes to existing DT infrastructure. (Galdino et al., 2025), in their requirement analysis for LLM-based cognitive assistants in QMS, propose a framework of this type that aligns LLM capabilities with ISO 9001:2015 quality management process requirements.

The second architecture is the Cognitive Digital Twin model, in which the LLM is embedded as an integral cognitive component within the DT framework, operating in close coupling with the DT's simulation, analytics, and data management layers. As mentioned in (Lou et al., 2025) in terms of five-dimensional cognitive DT approach and (Warburton, 2025) on hyper-adaptive production system architecture, this proposed architecture suggests LLM agent monitors the DT state persistently, proactively creates quality insight, and coordinates adaptive actions without any further requests from humans. Cognitive DT represents an autonomous quality management agent which resides in the DT Cyber-Physical system that makes dynamic real-time quality decisions within specific governance boundaries.

It is within this Cognitive Digital Twin paradigm that the LLM-DT-SQM framework proposed in this thesis, and fully discussed in *Chapter 3*, is situated, with tailored adaptations to address the specific requirements of strategic quality management in smart factory environments.

The third architecture is the Multi-Agent LLM-DT Orchestration model, characterized by a distributed array of specialized LLM agents, each responsible for a specific quality management domain (process monitoring, root-cause analysis, compliance documentation,

supplier quality management), coordinated by a central orchestrator agent that routes queries, manages information flows between agents, and synthesizes outputs into unified quality management responses. This architecture, exemplified by the LUMEN system of (Videsjorden et al., 2025) and the manufacturing automation frameworks of (Lim et al., 2024; Xia et al., 2025), provides the scalability and specialization required for enterprise-level DT-SQM deployments in complex manufacturing environments.

2.6 LLM-Driven Strategic Quality Management

LLM integration endows DT-SQM platforms with quality management capabilities that were previously impractical at operational scale. *Table 5* summarizes the principal capability domains and their evidence base before the two highest-value applications, natural language quality auditing and automated RCA, are examined in depth.

Table 5 - LLM Capabilities and their Evidence-Based SQM Applications

LLM Capability	SQM Application	Evidence / Source
Natural language understanding	Interactive quality audits; NL querying of DT data	Galdino et al. (2025); Yao et al. (2024): 24% audit efficiency gain
Multi-step reasoning	Automated root-cause analysis; failure chain tracing	Gill et al. (2025); Zhang et al. (2025): fault diagnosis up to 50% faster
Knowledge graph integration	Domain-specific quality reasoning; regulatory compliance	Meng et al. (2025); Videsjorden et al. (2025)
In-context learning	Zero-shot adaptation to new quality scenarios	Lin et al. (2024); Lou et al. (2025)
Report generation	Automated CAPA reports; audit documentation	Chowdhury (2025): 30% MES setup time reduction

2.6.1 Natural Language Querying for Real-Time Quality Audits

The application of LLM-powered natural language querying to real-time quality audits represents one of the most immediately deployable and organizationally impactful use cases for LLM-DT-SQM integration. The process of quality auditing involves analyzing process outputs, such as product evidence and records, with the identified processes to meet quality standard requirements. A substantial investment of time is required for quality auditing; specialists who undertake the process perform it, ordinarily on an intermittent basis, where they need to systematically identify and merge product or record results with standard requirements. The results are generated by manually extracting data from different systems and then comparison to the documented standards is undertaken (Galdino et al., 2025; Yao et al., 2024).

The DT-SQM with LLMs integrates naturally, creating systems capable of performing automatic quality auditing. Quality managers can pose audit-style queries, 'Is the tempering process on Line 2 currently operating within its ISO 9001-compliant control plan limits?' or 'Show me all non-conformances on supplier X's components in the last 30 days and summarize the root-cause categories', and receive comprehensive, evidence-grounded responses that draw on real-time DT data, historical quality records, and the relevant quality standard requirements simultaneously. Through structured API calls to the DT data services, QMS data bases and knowledge databases the ability of LLM to combine these different data sources leads to producing audit quality answers in a matter of seconds, not days. (Chowdhury, 2025; Yao et al., 2024).

The empirical evidence to support such a capability can be illustrated with (Yao et al., 2024), a smart auditing system enabled by LLMs which resulted in a 24% more efficient audit process as against manual processes by dynamically assessing and analyzing risks and compliance based on LLM inputs. From a Strategic Quality Management point of view, the copilot architecture in the form of the LLM assisting, but not substituting the human auditor, is significant as it retains human ownership and judgement for quality governance, yet significantly relieves the data retrieval and analytic burden of the human auditor such that more frequent, deeper audits are possible within a given organizational constraint.

The integration of LLM-powered auditing with Digital Twin simulation capabilities extends this value further, enabling prospective audit scenarios in addition to retrospective reviews. By querying the DT's simulation layer, an LLM-enhanced audit system can assess not only whether current processes comply with quality requirements but also whether planned process changes are likely to maintain compliance, supporting proactive quality governance and risk management (Galdino et al., 2025; Xia et al., 2025).

2.6.2 Automated Root Cause Analysis (RCA) and Corrective Action Reports

Root Cause Analysis (RCA) represents one of the most cognitively demanding and time-consuming activities in industrial SQM, requiring the quality engineer to trace a quality deviation or failure event through a complex, multi-dimensional causal chain, considering contributing process parameters, equipment conditions, material properties, operator behaviors, and environmental factors, to identify the fundamental cause and recommend targeted corrective actions (Gill et al., 2025; Zhang et al., 2025). Traditional RCA methods, including the 5 Whys, Fishbone (Ishikawa) diagrams, and Failure Mode and Effects Analysis (FMEA), provide structured analytical frameworks but rely heavily on expert judgment.

The development of an LLM based DT systems enables automation in much of the RCA pipeline by taking advantage of the multi-step causal reasoning abilities of the LLM on time-series data obtained from the DT as well as stored quality history and domain-based knowledge resources. The automated RCA process in such a system proceeds through several analytically structured phases: first, the LLM retrieves the relevant time-window sensor data from the DT layer surrounding the quality event; second, it identifies parameters that deviated from normal

operating ranges using statistical analysis functions invoked via API; third, it cross-references these deviations with the knowledge graph's causal model of process-quality relationships to identify likely contributing factors; fourth, it retrieves relevant historical root-cause analyses for similar events to contextualize the current findings; and fifth, it generates a structured, human-readable Corrective and Preventive Action (CAPA) report that documents the identified root cause, supporting evidence, recommended actions, and relevant quality standard clauses (Gill et al., 2025; Meng et al., 2025; Zhang et al., 2025).

(Gill et al., 2025) provided compelling empirical evidence for the feasibility of this approach in their work on LLM agent-based fault handling in process plants. Their system, which connected LLM agents to a DT simulation of a chemical process facility, demonstrated the autonomous generation of fault diagnosis reports and corrective action plans with minimal human prompting, effectively handling complex multi-cause fault scenarios that would previously have required extended expert analysis. (Zhang et al., 2025) reported complementary findings in a CNC machining quality DT context, where a full-stack platform integrating IoT data, cloud analytics, knowledge graphs, and LLM reasoning achieved Bayesian anomaly detection and automated root-cause attribution for tool wear and dimensional quality deviations, resulting in measurable reductions in machining errors and downtime.

2.7 State-of-the-Art: Early Applications of LLM-Enhanced DTs

The earliest architecturally coherent LLM-DT integrations, appearing in the 2023 literature, focused on establishing the foundational connection between LLM agents and DT simulation interfaces for manufacturing automation. (Xia et al., 2023) introduced a flexible modular production system enhanced with LLM agents capable of autonomously controlling atomic manufacturing functionalities, initiating, monitoring, and adapting production operations through natural language instructions interpreted by the LLM and translated into API calls to the DT layer. This work established the proof of concept that LLMs could function as semantic controllers for DT-integrated industrial automation, without requiring explicit programming for each manufacturing scenario.

In the 2024 literature, applications expanded significantly in scope and sophistication. (Xia et al., 2024) developed a multi-agent LLM system specifically designed for simulation model parametrization in DTs, demonstrating that LLM agents could autonomously configure complex DT simulation models through heuristic reasoning and iterative API interaction, reducing the configuration time and specialist expertise requirements for DT deployment. (Lin et al., 2024) proposed a Dynamic Data-Driven Generative Digital Twin (DDD-GenDT) framework that leveraged generative AI capabilities to enable zero-shot prediction of machining behavior, adaptively generating DT models for new process configurations without retraining, addressing one of the most significant scalability barriers in traditional DT deployment.

The quality-specific application domain gained traction in the 2024–2025 period. (Yao et al., 2024) demonstrated the 24% audit efficiency improvement through LLM-powered smart auditing referenced in *Section 2.6.1*. (J. Li et al., 2025) developed ChatSync, an LLM-powered

spatial-temporal knowledge reasoning system for production logistics synchronization, achieving 91.2% delivery punctuality through chat-based resource allocation and production traceability, illustrating the operational quality management value of LLM-IoT integration beyond the traditional quality control focus. (Chowdhury, 2025) reported a 30% reduction in MES setup time through an LLM-driven configuration recommendation system that applied generative AI to manufacturing execution system optimization.

Table 6 provides a structured comparative overview of the key early LLM-enhanced DT applications identified in the literature, organized by application domain, architectural pattern, and reported quality or efficiency outcome.

Table 6 - State-of-the-Art LLM-Enhanced Digital Twin Applications

Study	Application	Architecture	Key Result
Lou et al. (2025)	Cognitive DT for I5.0	Five-dimensional LLM agent + DT	Adaptive real-time incident handling
Xia et al. (2025)	LLM-DT-Automation integration	Three-layer cognitive–bridge–physical	Flexible automation control; simulation parametrization
Gill et al. (2025)	Fault handling in process plants	LLM agents + DT simulation	Autonomous CAPA with minimal reprompts
Zhang et al. (2025)	CNC machining quality DT	IoT + knowledge graph + LLM + cloud	Reduced tool wear; Bayesian anomaly detection
Yao et al. (2024)	Supplier quality audit system	LLM compliance copilot + DT risk model	24% audit process efficiency improvement
Videsjorden et al. (2025)	IoT system observability (LUMEN)	Multi-agent LLM + knowledge graphs	Semantic reasoning; enhanced system adaptability
Hossain & Hossain (2025)	Lean Six Sigma 4.0 + AI DT	LSS + AI + DT performance framework	Significant OEE and MTBF improvements

Four key conclusions may be drawn from this state-of-the-art investigation. First, the field is characterized by rapid technical progress but limited standardization: each of the applications reviewed employs a distinct architectural pattern, API integration approach, and evaluation methodology, making cross-study comparison and generalization challenging. Second, the application portfolio is weighted toward technical feasibility demonstrations and proof-of-concept implementations, with limited systematic evaluation of organizational adoption factors, scalability to industrial production environments, and long-term operational sustainability. Third, and most significantly for the purposes of this thesis, the existing

applications are predominantly oriented toward general manufacturing automation and fault management rather than the specific requirements of Strategic Quality Management, including quality standard compliance, audit trail documentation, CAPA management, and quality KPI governance. This gap defines the original research contribution that the proposed LLM-DT-SQM framework in *Chapter 3* is designed to address.

2.8 Security, Ethics, and Latency Constraints

The integration of Large Language Models into IoT-enabled Digital Twin frameworks for Strategic Quality Management introduces a set of critical challenges that extend beyond the technical architecture into the domains of security, ethics, regulatory compliance, and human factors. These limitations do not negate the fact that LLM-DT-SQM integration is a potential transformational opportunity but indicate that these are boundary conditions which need to be met before we can use LLM-DT-SQM integration in a manner that is responsible, reliable and sustainable. Four interdependent challenge areas will be investigated in this section: Privacy considerations for cloud LLM frameworks, deterministicity issues for non-deterministic AI, issues of morality/ethics of AI powered quality governance and the problem of the human-computer skills gap during LLM integration.

Data Privacy in Cloud-Based LLM Architectures

The dominant deployment model for state-of-the-art LLMs, cloud-based API access to models such as GPT-4, Claude, and Gemini, introduces significant data privacy risks in industrial manufacturing contexts that are not present in traditional DT-IoT deployments. When quality-critical manufacturing data, including process parameters, product specifications, supplier information, and non-conformance records, is transmitted to external LLM API endpoints for processing, it potentially exposes commercially sensitive and regulated information to cloud service providers and, in the event of security breaches, to unauthorized third parties (Bieniek et al., 2025; Doğruluk & Acikgoz, 2025; Y. Li et al., 2023).

Some technical mechanisms can be applied to resolve cloud data privacy issues of the LLM-DT-SQM system architecture. Federated Learning to allow the adaptation and finetuning of the LLM model using the on-premises manufactured data distributed across the factory without directly transmitting raw data outside of the boundary so that the data sovereignty is well preserved but the performance could still be improved over specialized task as specific data quality management purposes. On-premise deployment the LLMs models. We can utilize some readily available open source LLM, like LLaMA-3, Mistral or specifically finetuned industrial level version to be trained completely on premise by not transmitting any data outside to the cloud but this requires a substantial investment in computational resources and it should have relative worse performance than those frontier LLM on the cloud. Data anonymization and differential privacy techniques can reduce the sensitivity of data transmitted to cloud LLMs, though care must be taken to ensure that the anonymization does not remove the contextual specificity required for accurate quality reasoning (Bieniek et al., 2025; Doğruluk & Acikgoz, 2025).

Ensuring Determinism in Non-Deterministic AI Systems

A fundamental tension at the heart of LLM-DT-SQM integration is the contrast between the deterministic, quantitatively precise requirements of industrial quality management, where process parameters must be controlled to defined tolerances, quality decisions must be reproducible and auditable, and CAPA reports must comply with documented quality standard requirements, and the inherently non-deterministic, probabilistic nature of LLM inference (Galdino et al., 2025; Xia et al., 2025).

LLMs generate outputs by sampling from probability distributions over token sequences, meaning that identical queries may produce different responses on different invocations, and that LLMs can generate plausible sounding but factually incorrect statements, a phenomenon termed 'hallucination', particularly when operating at the boundary of their training distribution or in the absence of relevant contextual information. In a quality management context, hallucinated outputs could manifest as incorrect root-cause attributions, erroneous compliance assessments, or fabricated quality records, any of which could have serious consequences for product safety, regulatory compliance, and organizational liability (Galdino et al., 2025; Xia et al., 2025).

Several architectural strategies are available to constrain LLM non-determinism in SQM applications. Temperature control, setting the LLM's sampling temperature to low values, reduces output variability and encourages more conservative, factually grounded responses at the cost of reduced creative and synthetic reasoning capacity. Retrieval-Augmented Generation (RAG) grounds LLM responses in verified, retrieved information rather than purely parametric knowledge, reducing hallucination by anchoring outputs to documented facts (Lou et al., 2025; Meng et al., 2025). Human-in-the-loop validation architectures ensure that LLM-generated quality assessments and CAPA recommendations are reviewed and approved by qualified quality engineers before being recorded in the quality management system, preserving the accountability and traceability requirements of formal quality governance (Galdino et al., 2025; Xia et al., 2025). Hybrid architectures that combine LLM reasoning with rule-based quality logic, where the LLM handles natural language processing and synthesis while deterministic algorithms handle numerical computations and threshold comparisons, are emerging as a particularly promising approach to balancing flexibility and reliability (Hossain et al., 2025; Y. Li et al., 2024).

The Skills Gap: Human Factors in LLM Adoption

The successful adoption of LLM-enhanced DT-SQM systems is not solely a technical challenge; it is equally an organizational and human factors challenge. The shift from dashboard-based to conversational intelligence interfaces requires quality professionals to develop new cognitive and operational skills, adapt their workflows, and trust the analytical outputs of AI systems in domains where they have traditionally relied on their own specialized expertise (Galdino et al., 2025; Ouerghemmi & Ertz, 2025).

Empirical evidence from the reviewed literature suggests that resistance to LLM adoption among manufacturing professionals is significant and persistent. (Freire et al., 2024) observed, in their user study of LLM-powered knowledge sharing in manufacturing, that participants demonstrated a strong preference for human expert consultation over LLM interaction even when the LLM achieved faster and comparably accurate responses, reflecting a trust deficit that is not readily resolved by demonstrating technical performance alone. (Kovalchuk, 2025) identified workforce adaptation challenges as one of the primary barriers to AI integration in quality management, alongside technical and cybersecurity barriers, and recommended human-centered implementation strategies that emphasize co-design with quality teams, transparent explanation of AI decision processes, and phased deployment with progressive expansion of LLM autonomy as organizational trust develops.

Table 7 provides a consolidated summary of the key security, ethics, and constraints challenges for LLM-DT-SQM systems, together with the proposed mitigation approaches identified in the reviewed literature.

Table 7 - Security, Ethics, and Operational Constraints

Challenge Domain	Key Issues	Proposed Mitigations
Data Privacy & Security	GDPR/IEC 62443 compliance; IP protection; cloud data exposure	Federated learning; homomorphic encryption; on-premises LLM deployment
Latency Constraints	Sub-10ms requirements for real-time control on legacy PLCs; edge resource limits	Model compression (quantization, pruning); edge AI co-processors; async inference
Non-Determinism / Hallucination	LLM outputs are probabilistic; unreliable for safety-critical exact computations	Hybrid rule-based validation; human-in-the-loop guardrails; RAG grounding
Ethical Accountability	Opaque AI decisions; explainability gaps; bias in quality data	Explainable AI (XAI); audit trails; ISO 42001 AI governance
Human Factors & Skills Gap	Resistance to change; insufficient AI literacy; over-reliance on LLM outputs	Human-centered UX design; upskilling programs; phased adoption roadmaps

The challenges documented in this section, collectively define the critical boundary conditions for the responsible deployment of LLM-DT-SQM integration in industrial environments.

This agenda provides the foundation for the use case and experimental evaluation presented in *Chapters 3* and *4*.

3 LLM-DT-SQM Framework

3.1 Purpose and Role of the LLM-DT Hybrid Architecture

The proposed framework addresses the fragmented, data-rich, and knowledge-intensive nature of SQM in Smart Factories. It directly targets the five research gaps identified in *Section 2.4.* Standalone digital twin or predictive analytics approaches often struggle to capture this complexity while remaining flexible enough for human-in-the-loop governance.

The use of an LLM is motivated by the fact that quality management requires structured, context-dependent reasoning over multi-modal data, process logs, inspection results, standard clauses, historical failure modes, that is difficult to capture exhaustively with fixed heuristics or traditional supervised learning. Unlike rule-based systems, the LLM leverages generalized domain knowledge to interpret natural language queries, generate root-cause narratives, and propose corrective actions directly from heterogeneous quality data. Importantly, in this work the LLM is not used as an autonomous controller, but as a ***proposal and explanation generator*** operating under strict deterministic validation (Wang et al., 2024; Xia et al., 2025).

In the proposed architecture, the responsibility for maintaining a high-fidelity representation of the process, detecting anomalies, and simulating corrective scenarios remains with the digital twin. Each possible action is checked against a quality function that varies for each specific problem and includes constraints such as tolerances (dimensions, surface quality) or regulatory constraints. The DT is allowed to perform continuous state-verification, process-simulation and what-if analysis, whereas the LLM offers three capabilities complementary to each other: explanation and understandable summary of complex quality information and a reasoning engine providing step-by-step causal information for the purposes of root-cause analysis; a mechanism that provides man-friendly quality governance: questioning the LLM in natural language and handling audit-related tasks. In the review phase, the quality engineer may accept, modify, or override proposed actions; all such interventions are validated against the DT's deterministic model to ensure compliance with mandatory structural constraints before deployment (Khdoudi et al., 2024; S. Liu et al., 2023).

The objective of this approach is not to replace conventional quality assurance techniques with a language model, but to integrate the LLM into the governance loop as a flexible, high-level reasoning layer. The digital twin provides deterministic, auditable state representation, simulation, and constraint enforcement, while the LLM adds interpretability, adaptability, and richer heuristic expression to the quality management process. The next section describes the organization of these components in a multi-level architecture that coordinates data flow,

simulation models, generative reasoning, and operator interactions.

3.2 Conceptual Architecture: LLM-DT-SQM Framework

The proposed LLM-DT-SQM framework constitutes a multi-layered hybrid architecture that systematically integrates IoT-enabled Digital Twins (DTs) with Large Language Model (LLM) capabilities to support Strategic Quality Management (SQM) in Smart Factory environments. The framework is conceived as a closed-loop, human-supervised system in which real-time physical data, semantic knowledge, and generative AI reasoning are coordinated across five hierarchically ordered functional layers: the Input Layer, the Digital Twin & Knowledge Layer, the Generative AI Layer, the Human Interaction Layer, and the Data Management Layer. These layers are supported by a transversal API & Integration Backbone, described in Section 3.3, which provides the unified communication infrastructure shared by all layers.

The structure is represented in *Figure 5* and summarized in *Table 9*, which specifies the principal role, inputs, and outputs of each layer.

Table 8 makes this correspondence explicit, mapping each of the five research gaps identified in *Chapter 2* to the specific framework component that resolves it, and documenting the architectural rationale for that design choice.

Table 8 - Explicit mapping of the five research gaps to the framework components

Research gap	Framework response	Primary component	Architectural rationale
Interface & Interpretability	- Natural language audit interface - NLU/NLG pipeline - DT-grounded responses in < 3 s	- Generative AI Layer - Human Interaction Layer, Conversational Audit Interface	An LLM interface layer eliminates the specialist mediation currently required to extract insight from DT dashboards. Quality engineers query in natural language; the system retrieves, reasons, and responds.
CPS Fragmentation	- Multi-protocol adapter stack - Unified API Gateway - Data Quality Gateway	- API & Integration Backbone - Input Layer, Data Quality Gateway	The Backbone's protocol adapters translate OPC-UA, MQTT, PROFINET, and REST into a unified internal representation, decoupling functional layers from infrastructure heterogeneity.
Absent Reasoning Layer	- 6-stage LLM cognitive pipeline - Knowledge Graph causal ontology - RAG context assembly	- Generative AI Layer, Causal Reasoning stage - Digital Twin & Knowledge Layer, Data & Knowledge sub-layer	The Knowledge Graph provides the formal causal structure over which the LLM reasons; the pipeline formalizes RCA as a multi-hop ontology traversal rather than a free-form generation task.
Knowledge Accessibility	- Conversational Decision Support - RCA/PdM-Quality Viewer - KPI Dashboard	- Human Interaction Layer, SQM Application - Generative AI Layer, NLG	The SQM Application presents DT-derived quality intelligence through role-adapted interfaces. The LLM generates explanations, not just data, removing the software expertise barrier.
Standardization Deficit	- Five-layer closed-loop architecture - Five-function API catalog - Human governance gates	- The LLM-DT-SQM framework, the first unified architectural specification for this integration domain	No unified framework existed for LLM-DT-SQM integration. This work provides a replicable, fully specified architecture with defined interfaces, governance mechanisms, and ISO compliance mappings.

3.2.1 Justification of Key Design Choices

Three architectural decisions warrant explicit justification before the layer-by-layer description, as they represent deliberate choices among design alternatives.

The first decision is the five-layer hierarchical structure rather than a flat event-driven architecture. The flat structure has the advantage of lowering latencies for the data flows. However, there would be a direct tight binding between the LLM inference engine and the individual IIoT sensor data flows, making system's determinism Quality control reliant upon the availability of the LLM and adding non-deterministic delays to the control logic that are safety relevant. Hierarchal layer model does directly decouple the actual real-time physical monitoring (Input, Digital Twin & Knowledge layers with < 1s delay) from logical reasoning (Generative AI layer with 2-4s delay), maintaining the deterministic nature of the shopfloor control operations regardless of the LLM inference time. This separation is consistent with the tiered control architecture validated by (Bakhshandeh et al., 2024; Xia et al., 2025).

The second decision is the six-stage LLM cognitive pipeline with an automated Validation Gate rather than a simpler prompt-and-respond pattern. Each stage addresses a specific failure mode: RAG context assembly (Stage 2) prevents hallucination of CAPA references by grounding reasoning in verified historical records; the Validation Gate (Stage 6) prevents schema errors and numerically implausible outputs from reaching human reviewers. Removing either stage would reduce computational overhead but would transfer error-filtering responsibility to the human reviewer, increasing cognitive load and undermining the efficiency gains that motivate LLM adoption. The pipeline structure also makes each reasoning step independently auditable, satisfying the traceability requirements of ISO 9001:2015 Clause 7.5.3 (Lou et al., 2025; Meng et al., 2025).

The third decision is the use of two governance gates rather than one. Gate 1 (LLM Output Valid?) performs automated technical validation; it filters outputs that are numerically implausible or structurally malformed before they consume human attention. Gate 2 (Action Approved?) enforces human authority over consequential decisions; it cannot be automated without transferring accountability for production system changes from a qualified human to an AI agent. A single gate combining both functions would either automate decisions that require human accountability or require human review of technically invalid outputs. The two-gate architecture keeps automated checks separate from human judgment, which is the minimum governance structure consistent with ISO 9001:2015 Clause 8.5.2 (Galdino et al., 2025; Xia et al., 2025).

3.2.2 Framework Overview

The following *Table 9* summarizes the five layers and the API & Integration Backbone, specifying the role, inputs, and outputs of each component.

Figure 5 presents the complete framework as a visual diagram.

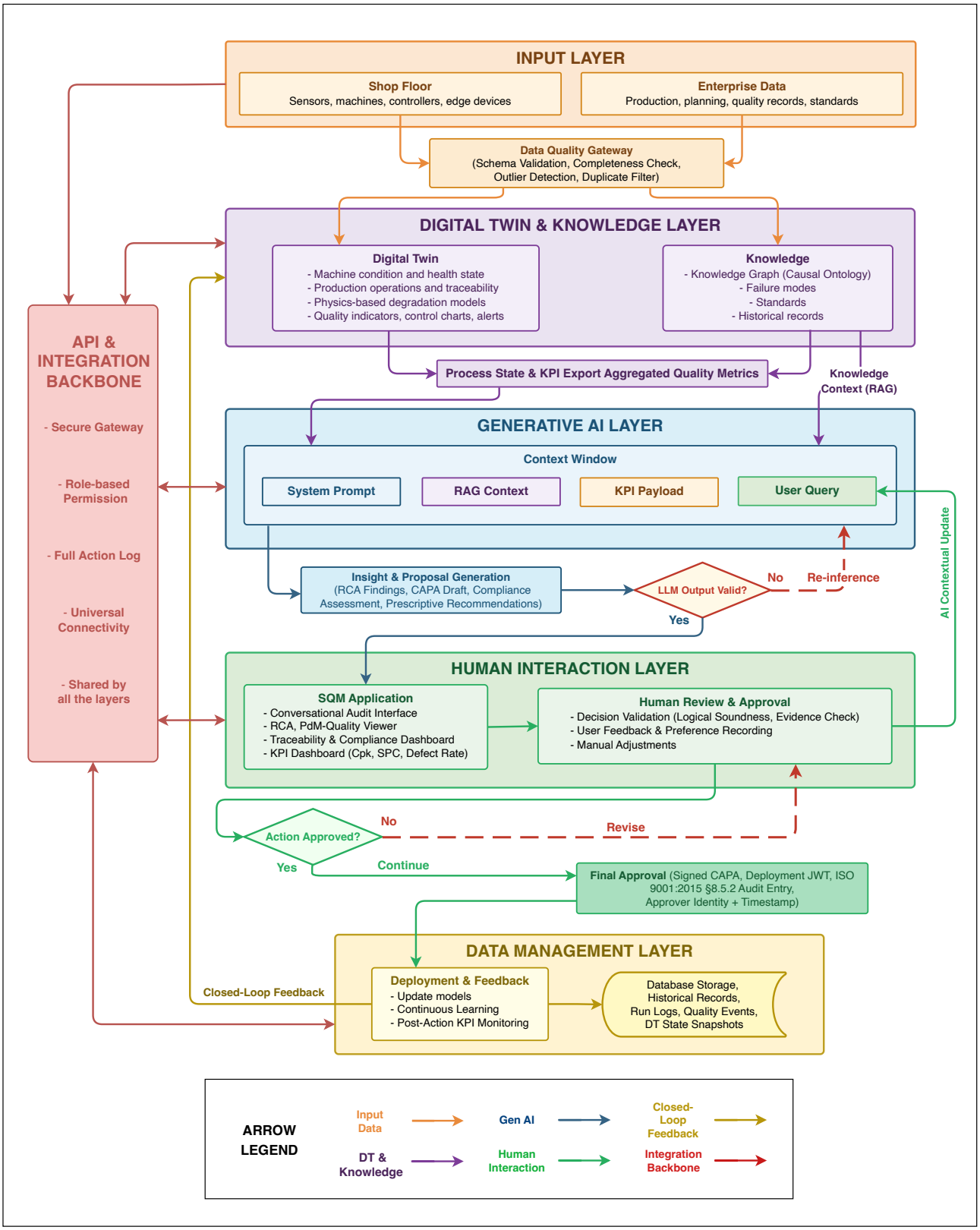

Figure 5 - LLM-DT-SQM framework

Table 9 - Functional summary of the five layers and the API & Integration Backbone.

Layer	Role	Principal inputs	Principal outputs
Input Layer	Aggregates, validates, and structures all operational and enterprise data required for quality governance.	- IIoT sensor streams (vibration, temperature, tool wear) - Machine state data (CNC, PLC, robots) - MES/ERP/PLM data (orders, specs, BOM) - Control plans and quality standards	- Unified time-series data feeds - Structured production context - Initial quality metadata for the Digital Twin
DT & Knowledge Layer	Synchronizes, simulates, and contextualizes data to build a high-fidelity representation of the process.	- Real-time sensor and machine data - Historical quality records - Process simulation parameters - Knowledge graph (failure modes, FMEA)	- Live Digital Twin process state - Predictive quality KPIs (drift, stability) - Process fingerprints and SPC charts - Knowledge graph-enriched context
Generative AI Layer	Performs semantic reasoning and generates actionable quality intelligence using the LLM cognitive layer.	- Digital Twin state and KPIs - Knowledge graph rules and standards - Natural language queries from users - Anomaly alerts and raw data patterns	- Automated root-cause analysis (RCA) - Draft CAPA reports and corrective actions - Audit summaries and compliance checks - Prescriptive quality recommendations
Human Interaction Layer	Supports review, validation, and final approval of AI-generated quality insights under human authority.	- RCA/CAPA proposals from the LLM - KPI dashboards and explanatory narratives - Raw quality data for cross-check - User feedback and adjustments	- Approved quality decisions and CAPAs - Human-validated process changes - Documented audit trails and justifications - Refined quality intelligence for deployment
Data Management Layer	Persists artifacts and orchestrates actions to ensure traceability, reproducibility, and closed-loop improvement.	- Approved quality decisions - CAPA records and logs - Updated process parameters - Digital Twin state snapshots	- Structured quality data archives - MES/PLC actuation commands - Audit logs and compliance reports - Feedback signals to the Digital Twin for continuous learning
API & Integration Backbone	Cross-cutting communication and security infrastructure shared by all layers	- Inter-layer communication requests - Enterprise system API calls - LLM function invocations - Actuation commands	- Authenticated inter-layer data flows - Role-based access enforcement - Timestamped API audit log - Protocol translation

3.3 API & Integration Backbone

The API & Integration Backbone covers a transversal technical aspect within the scope of the overall LLM-DT-SQM. All the functional layers described above communicate with each other and with external entities such as IoT devices, enterprise systems and knowledge repositories over the API & Integration Backbone. In this respect, while each functional layer has a unique and singular purpose within the quality governance cycle (data collection, simulation, function generation, human interaction or data storage) the API & Integration Backbone, the API & Integration Backbone only connects the other layers and manages how the services access functionalities within each other (no quality intelligence is collected, created or maintained by the API & Integration Backbone itself) in terms of invocation rules, data formats, and security measures.

This architectural positioning reflects a well-established principle in service-oriented architecture: integration infrastructure must be explicitly separated from business logic (Manwatkar, 2025; Xia et al., 2025). Placing the Backbone in one single functional layer would force a "false-functional dependency" and hide its cross-cutting nature. The Backbone is equally important for: data fetching of Input layer, simulation calls of the Digital Twin & Knowledge layer, function calling of the Generative AI layer, action authorization in Human Interaction layer, and logging and response actions of Data Management layer.

The Backbone comprises four principal components, corresponding to the four labels shown in *Figure 5*: (i) the Secure Gateway, providing a single, standardized entry point for all programmatic interactions with consistent authentication, rate limiting, and payload validation; (ii) Role-Based Access Control, ensuring that each framework component is authorized only to invoke functions within its designated operational scope; (iii) the Action Log, time-stamping every API call and generating a comprehensive record of all inter-component interactions; and (iv) Universal Connectivity, providing the multi-protocol adapter stack that mediates between the heterogeneous communication standards employed across the Smart Factory ecosystem without requiring changes to existing infrastructure. The Backbone's access control layer ensures that LLM-generated actions with real-world operational consequences can only be executed after passing through both formal decision gates, with every actuation command carrying a verified authorization token generated at the Final Approval step (Bieniek et al., 2025; Doğruluk & Acikgoz, 2025).

3.3.1 Communication Pattern: Unidirectional and Bidirectional Connections

As illustrated in *Figure 5*, the five framework layers connect to the Backbone through two distinct communication patterns, unidirectional and bidirectional, that reflect the fundamentally different operational roles of each layer within the quality governance cycle.

The Input Layer has a unidirectional connection to the Backbone, from the Input Layer to the Backbone. This is because the input Layer solely provides a role in gathering and validating data and submitting it to the framework from outside world, shop floor equipment's and enterprise systems. As a counterpart, the input layer does not receive any instructions, configuration updates and reply signals to the system, to prevent to be an actuator with the support of LLM-DT-SQM framework, on shop floor sensors, controllers or an enterprise system. A unidirectional connection communicates this asymmetry clearly: data flows in, nothing flows back out through this layer. Representing this connection as bidirectional would incorrectly suggest that the framework issues commands to its data sources, which would misrepresent the operational boundary between the framework and the physical production infrastructure it monitors.

The remaining four layers, the Digital Twin & Knowledge Layer, the Generative AI Layer, the Human Interaction Layer, and the Data Management Layer, each connect to the Backbone through bidirectional connections. This corresponds to their bilateral exchanges with the Backbone and with each other. While the Digital Twin & Knowledge Layer simultaneously receives valid data payloads from the Input Layer and sends out the KPI Export Payload to the Generative AI Layer via the Backbone, the latter alternately sends the API function calls to trigger new data collection, simulation executions or creation of a quality report and receives the results back with the response payload. The Human Interaction Layer takes over the validated LLM output as input, and sends the quality reviewer's approval and authorization token back to the Backbone in order to authorize the final deployment of the new quality record, while the Data Management Layer first carries out and stores the approved quality actions and then triggers a new learning phase on the Digital Twin & Knowledge Layer through the "Continuous Learning Trigger" signal. Indeed, in every instance, this bilateral aspect defines the relationship between the individual layer and those surrounding it, in terms of exchange of command and reply, request and result, approval and acceptance which are necessary for the performance of the layer in the process of quality governance.

Table 10 above summarizes the connection pattern between the Integration Backbone and each framework layer, specifying whether the relationship is unidirectional or bidirectional, the architectural rationale for that choice, and the specific data flows in each direction.

Table 10 - Backbone connection pattern per layer: unidirectional vs. bidirectional

Layer	Connection	Direction rationale	Data flows
Input Layer	**Unidirectional**	Data provider only. The layer collects and validates data from external sources. It does not receive commands or actuation signals from the framework.	- **Outbound**: validated sensor and enterprise data payloads forwarded to the Backbone for downstream processing
Digital Twin & Knowledge Layer	**Bidirectional**	Receives input data and sends KPI Export Payload. Also receives Continuous Learning Trigger updates from the Data Management Layer.	- **Inbound**: validated data from Input Layer, model update signals from Data Management Layer - **Outbound**: KPI Export Payload to Generative AI Layer
Generative AI Layer	**Bidirectional**	Sends API function call requests (data retrieval, DT simulation, QMS record creation) and receives the corresponding results.	- **Inbound**: API function call responses (sensor history, simulation results, FMEA records, standard clauses) - **Outbound**: API function call requests, validated CAPA proposals to Human Interaction Layer
Human Interaction Layer	**Bidirectional**	Receives LLM output for presentation and sends the reviewer's approval decision and authorization token back to trigger deployment.	- **Inbound**: validated LLM proposals for reviewer presentation - **Outbound**: approval decision, authorization token, signed CAPA record to Data Management Layer
Data Management Layer	**Bidirectional**	Receives approved actions for execution and storage and sends learning feedback signals back to the Digital Twin & Knowledge Layer.	- **Inbound**: approved CAPA records, authorization tokens, quality event data - **Outbound**: actuation commands to production systems, Continuous Learning Trigger to Digital Twin & Knowledge Layer

3.4 Input Layer

The Input Layer represents the foundation of data collection & integration for the LLM-DT-SQM system. Its function is to bring together, validate and normalize all operational & enterprise data relevant to the quality governance. The layer itself consists of two sub-components, the Physical Layer and the Enterprise Data layer, acting in parallel which translate into the two sources of data powering the system, i.e. production data (operational)

and enterprise system data (enterprise), as per *Figure 5*. Both data flow then move to the Data Quality Gateway, where a series of 4 data validation rules were executed: Schema Validation (if payload matches definition), Completeness check (if there are missing compulsory data), Outlier check (statistical control of values outside normal range so to discard or inspect) & Duplicate filter. The framework cycle begins here, with the continuous collection and validation of these two streams before any downstream processing occurs (Bongomin et al., 2025; Christou et al., 2022).

3.4.1 Physical - Shop Floor

The Physical sub-layer encompasses the full range of IIoT-connected devices, sensors, and control systems that constitute the operational environment of the Smart Factory. This includes vibration, temperature, and tool-wear sensors attached to production equipment; machine state monitoring systems associated with CNC machining centers, industrial robots, and PLCs; and the edge computing infrastructure responsible for local data pre-processing, buffering, and protocol normalization. Communication between shop-floor devices and the upstream framework layers is mediated through the API & Integration Backbone's multi-protocol adapter stack (Christou et al., 2022; Monek & Fischer, 2023).

Step 1, Measurement of sensor data is performed with accelerometers for the vibration signals at the 1 kHz-10 kHz range, PT100 resistance thermometers or thermocouples are used to determine the temperature, current clamp sensors are used for measuring the current, and thus the load of the spindle and dimension probes for an in-process measurement of geometry. All the signals are buffering on the edge gateway and timestamp.

Step 2, Edge Pre-processing: Signals are filtered with a low-pass Butterworth filter of order 4, with a cut-off frequency of 400 Hz to remove spurious high-frequency components. They are scaled so the values lie in [0,1]. Timestamps are synchronized through NTP/PTP so that their maximum difference between instruments is less than 1 second. Anomalous readings are flagged using a Hampel identifier operating over a rolling window (window size = 7, threshold = 3σ), which suppresses isolated spike artefacts without distorting the underlying signal trajectory (Doğruluk & Acikgoz, 2025).

Step 3, Protocol Encoding: Processed data is encoded for the target protocol and forwarded to the API & Integration Backbone: CNC/PLC assets via OPC-UA NodeID mapping; IIoT sensors via MQTT (factory/line/cell/asset/param); safety PLCs via PROFINET below 1 ms.

Step 4, Data Quality Gateway: JSON Schema / OPC-UA Information Model validation, completeness checks, outlier detection, and duplicate filtering. Valid payloads are forwarded to the Digital Twin & Knowledge Layer; rejected payloads are logged with error codes.

The following Python *Code 4* excerpt illustrates the edge pre-processing and MQTT dispatch logic:

Code 4 - Python: Edge signal pre-processing and MQTT dispatch to the DT data broker.

```
# Python: Edge pre-processing pipeline (simplified)
import numpy as np
from scipy.signal import butter, filtfilt


def edge_preprocess(raw_signal, fs=1000):
    # Step 1: Butterworth low-pass filter (cutoff 400 Hz)
    b, a = butter(4, 400 / (0.5 * fs), btype='low')
    filtered = filtfilt(b, a, raw_signal)
    # Step 2: Normalize to [0, 1]
    norm = (filtered - filtered.min()) / (filtered.max() - filtered.min() + 1e-9)
    # Step 3: Hampel outlier detection (window=7, threshold=3σ)
    median = np.median(norm)
    mad = np.median(np.abs(norm - median))
    anomaly_mask = np.abs(norm - median) > 3 * 1.4826 * mad
    return {
        'signal': norm.tolist(),
        'anomaly_flags': anomaly_mask.tolist(),
        'quality_score': 1.0 - anomaly_mask.mean(),
        'timestamp_utc': get_ntp_timestamp()   # NTP-synced
    }


# Publish to Digital Twin broker via MQTT
client.publish(
    topic='factory/line1/cnc01/vibration/spindle',
    payload=json.dumps(edge_preprocess(raw_acc_data)),
    qos=1   # at-least-once delivery guarantee
)
```

The inclusion of edge computing within the Physical sub-layer is architecturally significant: edge nodes perform the initial data filtering, compression, and anomaly flagging, reducing the communication bandwidth required to transmit full-resolution sensor data to cloud or on-premises computation layers, and ensuring that latency constraints for real-time quality monitoring can be satisfied without compromising the safety-critical determinism of PLC-level control (Doğruluk & Acikgoz, 2025).

The *Table 11* below presents the protocol selection rationale for the Physical sub-layer, specifying the target asset types, communication latency requirements, and key technical characteristics of each protocol deployed within the LLM-DT-SQM framework.

Table 11 - Industrial Communication Protocol Selection for the Physical sub-layer

Protocol	Target Assets	Target Latency	Key Characteristics
OPC-UA	CNC machining centres, PLCs	< 100 ms	Semantic data modelling, security, NodeID mapping
MQTT	IIoT sensors, edge gateways	< 10 ms	Publish–subscribe, low bandwidth, QoS levels 0–2
PROFINET	Safety PLCs, motion controllers	< 1 ms	Hard real-time, deterministic frame scheduling
REST/HTTP	ERP, MES, cloud APIs	< 500 ms	Stateless, JSON/XML payloads, standard authentication

3.4.2 Enterprise Data

The Enterprise Data sub-layer covers the aggregated, structured data asset entities which form the quality governance context of the production operation. These would take in quality relevant data from Manufacturing Execution Systems (MES), Enterprise Resource Planning (ERP) systems, and Product Lifecycle Management (PLM) sources such as production orders, material bills-of-materials (BOMs), engineering specifications, control plans, quality norms and historical quality records. This layer provides the framework with the normative reference data, tolerances, acceptance criteria, process parameters, and regulatory requirements, against which real-time process states and LLM-generated quality assessments are evaluated (Sokolov et al., 2023; Tseng et al., 2025).

The integration of enterprise data with IIoT sensor streams within a unified Input Layer resolves a longstanding challenge in industrial quality management: the semantic gap between operational data (what the process is doing) and quality governance data (what the process should be doing). By consolidating these two categories of information at the framework's entry point, the LLM-DT-SQM architecture ensures that all downstream analytical and decision-support functions operate on a contextually complete data substrate, in which process states are always interpreted in relation to the applicable quality specifications and regulatory requirements (Y. Li et al., 2024).

3.5 Digital Twin & Knowledge Layer

The Digital Twin & Knowledge Layer receives the validated data streams from the Input Layer and performs the synchronization, simulation, and contextualization operations required to build a high-fidelity representation of the production process. It comprises two tightly coupled sub-components: the Digital Twin and the Knowledge base, whose outputs are

aggregated at the Process State & KPI Export step and forwarded to the Generative AI Layer (S. Liu et al., 2023; Sun et al., 2025).

3.5.1 Digital Twin

The Digital Twin sub-layer maintains the core cyber-physical representation of the Smart Factory's quality-relevant process states, continuously synchronized with the incoming data streams from the Input Layer. The DT's simulation capability, implemented through physics-informed process models and statistical learning, enables the evaluation of counterfactual scenarios and the prediction of quality outcomes under alternative process conditions (Khdoudi et al., 2024; Monek & Fischer, 2023).

The DT consists of four parallel operational capabilities as shown in *Figure 5*, corresponding to the four stages on the manufacturing quality causality chain: monitoring machine condition and health state: the DT monitors on line condition, operational mode, load and degrading tendency of production equipment and captures equipment dependent leading signal that explain the drift of product quality over time; monitoring production operations and traceability: the DT stores current production order and operation and tracks all manufacturing resources to achieve per-unit quality tracing form material consumption up to product consumption; physics based degradation modeling: by using mathematical model for characteristic physical phenomena which could degrade process quality due to machine degrading such as machine wear, temperature drift and mechanical fatigue, DT predicts quality influential variable from machine signals; computation of quality indicator, control charts and warning/alert: the DT integrates all these information to compute the main quality KPIs (e.g., yield), process capabilities index, statistical process control (e.g., limits and control status) and defect prediction at each monitoring quality feature and triggers an warning and alert.These four parallel capabilities construct a data chain and its output becomes input to the next stage's input; The data flow moves towards progressively precise result, from raw equipment data up to accurate and actionable product quality intelligence. (S. Liu et al., 2023; Xia et al., 2025).

The decision to structure the Digital Twin around four distinct operational capabilities rather than a monolithic model reflects the causal architecture of manufacturing quality. Quality outcomes in production processes arise from a chain of causal dependencies spanning four analytically distinct domains: equipment state, process execution, physical degradation phenomena, and product quality characteristics. A monolithic DT conflating these domains would produce a model that is simultaneously too coarse for physics-based defect prediction and too fine-grained for strategic quality KPI computation. The four-capability decomposition mirrors this causal chain explicitly: equipment condition data feeds the degradation models; degradation models feed the quality KPI computation; and quality KPI outputs drive the Generative AI Layer's reasoning. Each capability requires a distinct modelling paradigm and update frequency, real-time state tracking for machine condition, event-driven recording for process operations, model-based estimation for physical phenomena, and measurement-

integrated statistical computation for quality indicators. This architectural separation prevents modelling conflicts and enables each capability to be independently calibrated, updated, and validated without affecting the others (S. Liu et al., 2023).

Table 12 specifies for each Digital Twin capability: the monitored domain, the nature of inputs, the key outputs, the category of quality data produced, and the principal SQM function enabled.

Table 12 - Digital Twin four-capability architecture

DT capability	Inputs	Key outputs	Quality data category	SQM function
Machine condition & health state	Real-time equipment sensor signals Operational state data from controllers	Equipment health score Degradation trajectory	*Process condition data*, leading indicators of equipment-induced quality drift	Real-time asset status monitoring Input to physics-based degradation models
Production operations & traceability	Production order data from enterprise systems Machine event records (operation type, cycle time)	Per-operation process fingerprint Cycle time vs. standard deviation Measurement result per serial number	*Process execution data*, contextual traceability linking each operation to quality outcomes	Full per-unit production traceability Operation-level deviation attribution Regulatory traceability compliance (ISO 9001:2015 §8.5.2)
Physics-based degradation models	Machine condition outputs (health state, thermal state) Cumulative production load per asset	Degradation index per monitored phenomenon Predicted dimensional and quality deviation	*Physical phenomenon data*, intermediate predictors bridging equipment state and product quality deviation	Physics-coherent quality deviation prediction Pre-emptive maintenance scheduling before quality impact
Quality indicators, control charts, alerts	Degradation model outputs Process execution data In-process and end-of-process measurements	Cpk per quality characteristic SPC control chart status (Nelson rules)	*Product quality characteristic data*, lagging quality indicators	Primary KPI source for LLM context payload Real-time compliance checking vs. control plan Closed-loop model feedback after CAPA outcome

3.5.2 Knowledge

The Knowledge sub-layer provides the semantic and normative knowledge infrastructure that contextualizes the DT's numerical process representation within the broader quality governance framework. Its central component is a Knowledge Graph, implemented as an OWL/RDF causal graph encoding over 400 causal triples, that maps the relationships between process parameters, equipment conditions, material properties, and quality outcomes, as well as the applicable quality standards, failure mode libraries (FMEA), and CAPA templates that govern quality management decisions (Meng et al., 2025; Videsjorden et al., 2025).

As illustrated in *Figure 5*, the Knowledge sub-layer comprises four components, each corresponding to one of the bullet points in the Knowledge box of the framework diagram. The Knowledge Base Causal Ontology is a structured encoding of over 400 causal triples, in OWL/RDF, linking aspects such as process parameters, equipment states, material characteristics and output quality. Failure Modes contains the FMEA library storing records of the relevant failure modes and their impact, cause, detection, and RPN value. Standards comprises a representation of relevant ISO 9001:2015, AS9100 and IATF 16949:2016 requirements in JSON-LD format, allowing machine checks of compliance to clause level requirements.

Historical Records comprises a store of previously completed CARs/PARs as vector embeddings within a specialized vector database, allows access for RAG retrieval into an LLM based on semantic relevance. This collection of four components comprise the formal substrate of knowledge within the LLM, guiding causal reasoning.

The co-presence of the DT sub-layer and the Knowledge sub-layer within the Digital Twin & Knowledge Layer illustrate the architecture perspective where the effective quality intelligence relies on both quantity of the process description and quality of knowledge domain. Neither component alone is sufficient for strategic quality decision-making: numerical process data without normative context cannot be evaluated against quality specifications; domain knowledge without real-time process data cannot identify current deviations from quality requirements (Bongomin et al., 2025; Kabir et al., 2025).

3.5.3 Process State & KPI Export

The 'Process State & KPI Export' step, as shown in *Figure 5*, aggregates the Digital Twin's quality metrics, Cpk values, SPC Nelson rule violations, equipment health indicators, triggered alerts, with the Knowledge Context retrieved by the Knowledge sub-layer via RAG, and packages them as a structured JSON payload forwarded to the LLM Cognitive Layer via the Integration Backbone. This step defines a clear functional boundary between the physics-grounded Digital Twin & Knowledge Layer and the semantic reasoning functions of the Generative AI Layer, ensuring the LLM operates on pre-processed, contextualized quality intelligence rather than raw sensor data (Cui et al., 2023; Wang et al., 2024).

The following JSON excerpt, *Code 5*, illustrates the structure of the KPI export payload

forwarded to the LLM Context Window:

Code 5 - JSON: KPI Export Payload structure forwarded to the LLM Cognitive Layer

```
// JSON: KPI Export Payload, forwarded to LLM Cognitive Layer
{
  "timestamp": "2026-04-20T14:32:01Z",
  "asset_id": "CNC_LINE1_01",
  "quality_kpis": {
    "diameter_D1": { "cpk": 0.82, "mean": 49.97, "usl": 50.05, "lsl": 49.95,
                     "status": "WARNING" },
    "surface_Ra":  { "cpk": 1.34, "status": "OK" },
    "spc_nelson":  { "rule_2_triggered": true,
                     "description": "9 consecutive points above mean" }
  },
  "equipment_health": {
    "tool_wear_pct": 73.4,
    "vibration_rms": 4.82,
    "spindle_temp_delta": 12.3
  },
  "triggered_alerts": ["SPC_NELSON_R2", "TOOL_WEAR_THRESHOLD_70PCT"],
  "knowledge_context": {
    "relevant_fmea_ids": ["FM-CNC-014", "FM-CNC-022"],
    "similar_past_capas": ["CAPA-2025-118"]
  }
}
```

This export step is architecturally significant in that it defines a clear functional boundary between the physics-grounded, data-intensive Digital Twin & Knowledge Layer and the semantic reasoning functions of the Generative AI Layer, ensuring that the LLM operates on pre-processed, contextualized quality intelligence rather than raw sensor data.

3.6 Generative AI Layer

The Generative AI Layer constitutes the intellectual core of the LLM-DT-SQM framework. Its function, as specified in *Figure 5*, is to perform semantic reasoning over the quality intelligence received from the Digital Twin & Knowledge and to generate actionable quality management outputs, including root-cause analyses, corrective action proposals, compliance assessments, and natural language quality narratives. (Galdino et al., 2025; Ouerghemmi & Ertz, 2025).

3.6.1 LLM Cognitive Layer: Six-Stage Processing Pipeline

The LLM Cognitive Layer processes each quality inquiry or alert through a six-stage sequential pipeline, which is described below. This pipeline structure distinguishes the proposed framework from prior LLM-DT integrations by explicitly formalizing the context assembly, reasoning, and validation stages as discrete, auditable processing steps (Galdino et al., 2025; Ouerghemmi & Ertz, 2025).

Stage 1, Natural Language Understanding (NLU): Receives a user query or automated alert, performs an intent classification (audit / RCA / compliance / predictive) and a named

entity recognition (asset ID, prod line, time range, relevant quality standards). Based on intent and entities, the LLM creates an invocation plan for down stream tools (Freire et al., 2024).

Stage 2, RAG Context Assembly: A vector similarity search is executed against the CAPA Vector Store to retrieve the top-k semantically similar past cases. At same time, a query over Knowledge Graph through SPARQL would retrieve FMEAs against triggered alerts and the Standards Repository queried to get relevant ISO/AS/IATF clause text. The collected results are consolidated to construct a grounded context to the system prompt in order to ground the subsequent LLM reasoning on factual domain knowledge instead of parametric pre-training (Lou et al., 2025; Meng et al., 2025).

Stage 3, Causal Reasoning: Multi-step chain-of-thought reasoning is performed by the LLM across the KPI payload and recalled context; causal paths are traced from identified quality deviations through the contributing process parameters to the root cause equipment and material parameters. The causal ontology of the knowledge graph offers explicit relational structure that the LLM reasoning adheres to; causal inference paths must be causally valid and are limited to a parameterizable maximum inference chain length (hops) of four (Gill et al., 2025).

Stage 4, API Function Calling: Based on the analytical requirements identified during causal reasoning, the LLM autonomously invokes structured API functions, described in *Table 13*, to retrieve additional data through the API & Integration Backbone, execute DT simulations, or create quality records in enterprise systems. This stage may involve multiple sequential tool invocations (up to three rounds per query), with each tool's response enriching the LLM's analytical context for the subsequent reasoning step (Cui et al., 2023; Xia et al., 2025).

Stage 5, Insight & Proposal Generation: The LLM's Natural Language Generation (NLG) capability synthesizes the accumulated evidence into structured quality intelligence artefacts, including RCA narratives, CAPA draft reports with ISO clause references, compliance assessment summaries, and predictive quality forecasts formatted for presentation to human reviewers.

Stage 6, Validation Gate ('LLM Output Valid?'): An automated consistency and plausibility check is applied to the generated output prior to its presentation to human reviewers. Validation checks include numerical plausibility (KPI values within physical bounds), evidence citation verification (each RCA claim references at least one KPI data point), standard clause matching (compliance findings cite verifiable clause identifiers), hallucination filtering (CAPA references validated against actual QMS record IDs), and output schema compliance (JSON format matching the CAPA template specification). Outputs failing validation are returned to Stage 1 for re-inference with modified context parameters (Galdino et al., 2025; Xia et al., 2025).

Table 13 - LLM API Function Catalog

Function	Parameters	Returns	SQM Use Case
get_sensor_history()	asset_id, param, window	Time-series data array	Retrieve process data for RCA time window
run_dt_simulation()	scenario_config: {param, value}	Predicted quality outcome	Counterfactual scenario analysis for CDS
query_fmea()	failure_mode_keywords	Ranked FMEA entries + RPN	Causal chain grounding for LLM reasoning
get_spc_status()	characteristic_id	Nelson rules, Cpk, trend	Compliance checking and quality audit
create_ncr()	description, evidence, severity	NCR ticket ID in QMS	Automated non-conformance record creation
retrieve_standard_clause()	query, standard_id	Clause text + obligation type	Regulatory compliance verification and citation

The code excerpt in *Code 6*, illustrates the implementation of the RAG context assembly (Stage 2), the automated API function calling cycle (Stage 4), and the final output validation gate (Stage 6) of the proposed pipeline:

Code 6 - JavaScript: LLM function calling cycle and output validation gate, Stages 4–6

```
// JavaScript: Function calling cycle, Stage 4
const resp = await openai.chat.completions.create({
  model: 'gpt-4o', temperature: 0.1,  // Low temp for determinism
  messages: [{ role: 'system', content: systemPrompt },
             { role: 'user',   content: userQuery }],
  tools: FUNCTION_CATALOG, tool_choice: 'auto'
});
// Execute all tool calls returned by the LLM
if (resp.choices[0].finish_reason === 'tool_calls') {
  const results = await Promise.all(
    resp.choices[0].message.tool_calls.map(tc =>
      callApiFunction(tc.function.name, JSON.parse(tc.function.arguments))
    )
  );
  // Feed tool results back, LLM synthesizes final response (Stage 5)
  finalResp = await openai.chat.completions.create({
    messages: [...messages, ...results.map(r => ({ role: 'tool', content: r }))],
    temperature: 0.1
  });
}
// Validation Gate (Stage 6)
const valid = validateOutput(finalResp, kpiPayload);
if (!valid.passed) return { status: 'INVALID', retry: true };
```

3.6.2 Prompt Engineering and Context Window Design

The reliability and determinism of the LLM quality reasoning depend not only on the six-stage pipeline architecture described in *Section 3.6.1*, but on the deliberate design of the four-component context window assembled at each inference call. The first is the System Prompt, a fixed, versioned configuration block always fed to the LLM first, preceding any quality data and user prompt. It defines the LLM's operating persona as a quality engineering helper, domain-specific constraints it must observe-no parameter adjustments can occur outside the control plan, any corrective action can only stem from a FMEA entry, any compliance finding may only refer to a provable clause from an ISO/IATF standard-the correct JSON output, the chain-of-thought reasoning instructions, and the fixed inference temperature of 0.1 which gives (almost) non-deterministic output. By encoding these instructions as a static configuration rather than embedding them in each query, the System Prompt ensures governance constraints are persistent, auditable, and immune to injection from dynamically retrieved content (Galdino et al., 2025).

The complete context window presented to the LLM at each inference call is assembled from four ordered components, *Table 14*, each with a distinct informational role, as illustrated in *Figure 5*.

Table 14 - Context window assembly

#	Component	Content	Source	Role in reasoning
1	System Prompt *(instructions & rules)*	Role definition, domain constraints, output schema, reasoning protocol, temperature 0.1	Static framework configuration, version-controlled	Establishes operating context and governance constraints; placed first to orient all subsequent reasoning
2	RAG Context *(historical cases)*	Top-k similar CAPA records; relevant FMEA failure mode entries; applicable ISO/IATF clause text	CAPA Vector Store; FMEA Library; Standards Repository	Grounds reasoning in verified domain knowledge; prevents hallucination of FMEA or ISO references
3	KPI Payload *(live quality data)*	JSON from Quality Twin: Cpk values, SPC status, equipment health, triggered alerts, asset ID, timestamp	Quality Twin, real-time KPI export	Provides the specific numerical quality event; all output numerical claims must originate here
4	User Query *(engineer question)*	Natural language query from quality engineer, or structured alert from Quality Twin	Human Interaction Layer or Quality Twin alert	Specifies the required output type and the quality characteristic or time window of interest

The ordering of the four components is deliberate. The System Prompt is always first, establishing governance constraints before any domain-specific content is presented. The RAG Context follows, providing the normative and historical knowledge substrate before the specific numerical data. The KPI Payload comes third, presenting live quality event data within the normative context already established. The User Query is last, specifying the required output type after the LLM has been fully oriented. This sequence ensures the most likely possibility the LLM will use numerical input in the frame of reference of normative information and not the more loosely constrained frame, the core of critical reasoning appropriate to ISO normative quality governance. User Query comes from the Human Interaction Layer Conversational Audit Interface and traces back what has become the feedback path of the AI Contextual Update arrow in *Figure 5*. This path of information is different from the async and machine-driven Closed-Loop Feedback signal.

The color coding of the four context window components in *Figure 5* is not decorative, it explicitly communicates the layer of origin of each component, reflecting the cross-layer nature of the context window assembly. The RAG Context pill is colored to match the Digital Twin & Knowledge Layer because the retrieved historical cases, FMEA failure mode records, and ISO/IATF clause text are all drawn from the Knowledge sub-layer of that same layer, specifically from the CAPA Vector Store, the FMEA Library, and the Standards Repository described in Section 3.5.2. The KPI Payload pill is colored to match the Input Layer because it represents the aggregated result of the continuous data collection that begins there: the quality KPIs and alerts packaged by the Quality Twin are ultimately derived from the raw sensor and enterprise data that the Input Layer acquired, validated, and passed downstream. The color makes the traceability of this data lineage visually explicit, the LLM is reasoning over structured intelligence that originates, at its root, from the physical production environment. The User Query pill is colored to match the Human Interaction Layer because it is the direct expression of human intent in the quality governance cycle, submitted by the quality engineer through the Conversational Audit Interface. Its distinct color communicates that this component is the only element of the context window that is not system-generated, it carries the contextual judgment, operational awareness, and specific analytical need of the human professional who is accountable for the quality decision.

3.7 Human Interaction Layer

The Human Interaction Layer serves the critical governance function of ensuring that AI-generated quality intelligence remains subject to qualified human validation, adjustment, and final approval before it is deployed as an operative quality management action. This layer instantiates the human-in-the-loop governance principle that is essential for responsible AI adoption in quality-critical industrial environments, where the accountability and traceability requirements of quality standards such as ISO 9001:2015 cannot be delegated to autonomous AI systems without explicit human oversight (Galdino et al., 2025; Ouerghemmi & Ertz, 2025).

3.7.1 SQM Application

The SQM Application sub-layer presents the validated outputs of the Generative AI Layer to quality management practitioners through a suite of four quality-specific application interfaces: (i) the Conversational Audit Interface, enabling natural language queries with DT-grounded responses; (ii) the RCA/PdM-Quality Viewer, displaying root-cause findings alongside supporting DT process data and FMEA cross-references; (iii) the Traceability & Compliance Dashboard, visualizing quality event history and standard conformance status; and (iv) KPI Dashboards presenting predictive quality metrics in the context of applicable specification limits and control plan requirements (Freire et al., 2024; Yao et al., 2024).

The SQM Application is positioned in the Human Interaction Layer, rather than in the Generative AI Layer where the quality intelligence originates, for a fundamental architectural reason: presentation and decision support are human-facing functions that must be designed around the cognitive needs of quality professionals, not around the data structures of the AI system. The SQM Application translates the LLM's structured JSON outputs into role-appropriate visual representations that give the reviewer immediate access to the evidence chain, the confidence level, the FMEA cross-references, and the historical precedents that support or challenge the AI's recommendation. Without such a translation layer, the human reviewer would be faced with the direct outputs of the AI, and interpretation would necessarily rely on expert knowledge again making it impossible to satisfy the access principle of the system architecture and laying the groundwork for arbitrary approval. Here the SQM Application operationalizes the system architecture's human-in-the-loop principle and serves not just to display data, but to actively mediate the environment within which the human decision maker is positioned, facilitating confident and information-based approval (Freire et al., 2024; Yao et al., 2024).

The four interface components each address a distinct dimension of the quality governance decision. The Conversational Audit Interface removes the software expertise barrier identified in Gap 4 of the SLR, *Section 2.4*, enabling any quality professional to retrieve quality intelligence through natural language. The RCA/PdM-Quality Viewer provides the causal evidence needed to validate the LLM's root-cause attribution. The Traceability & Compliance Dashboard satisfies the documentation requirements of ISO 9001:2015 Clause 9.1 and IATF 16949:2016. The KPI Dashboard enables the reviewer to contextualize the current quality event within the broader process trend, confirming whether the AI's proposed action is consistent with the process history or whether additional context is required before approval.

3.7.2 Human Review & Approval

The Human Review & Approval sub-component provides the structured governance environment in which qualified quality professionals evaluate, validate, and if necessary, refine the LLM-generated quality proposals before authorizing their deployment. As shown in *Figure 5*, there are 3 activities for reviewing: Decision Validation (Logical Soundness, Evidence Check)

determines the factuality and logical consistency of the LLM's RCA and CAPA recommendation, by validating the attribution of the RCA to the underlying cause against reviewer's operational knowledge of the process. User Feedback & Preference Recording records the reviewer's annotations, corrections, and elaborations as structured signals for future system learning, expanding the CAPA Vector Store and informing the Knowledge Graph weight adjustment. Manual Adjustments allow the reviewer to directly modify the LLM's proposal, changing a parameter value, adding a failure mode not captured in the FMEA, or replacing a suggested corrective action, before the approval decision is made. Together, these three activities ensure that the review is substantive and evidence-grounded rather than a rubber-stamp approval (Galdino et al., 2025).

The AI Contextual Update arrow in *Figure 5*, the dashed feedback path returning from Human Review & Approval back to the LLM Cognitive Layer, represents one of the most important design choices in the framework. Here, if the reviewer has found a bug or a missing piece in the reasoning by the LLM, this path can add the reviewer's correction directly into the LLM context as an extra input and then kick off a new focused re-inference. This is distinctly different from simply starting pipeline anew in terms of the architecture: 'this failure mode is not in the FMEA', 'the coolant factor isn't mentioned in the RCA'; reviewer's annotation carries with it some "operational knowledge" to lead the second pass reasoning to just where it was lacking. The result is a collaborative dynamic in which the LLM's computational breadth and the reviewer's domain expertise are combined iteratively until the output is fit for approval. This pattern is consistent with the human-technology-organization interaction model described by (Galdino et al., 2025), who identify iterative human-AI co-reasoning as a primary deployment requirement for LLM-based quality management assistants.

In the standard review workflow, the reviewer first consults the RCA/PdM-Quality Viewer to inspect the LLM's causal reasoning chain and verify that the retrieved FMEA entries and analogous CAPA records support the proposed root-cause attribution. If the reviewer confirms the attribution in their field, the proposal moves on to Manual Adjustments where changing a parameter, such as setting a suggested process change from the proposed value to domain-specific conservative value, is tracked by a structured annotation. Such an annotation automatically triggers an AI Contextual Update path: when a domain annotation from the reviewer is provided, it's fed into the active context of the LLM as another input which the model iterates over to refine the output by incorporating domain-specific knowledge prior to the input's forwarding to Final Approval. This mechanism ensures that the Final Approval artefact reflects both the LLM's analytical breadth and the reviewer's operational knowledge, and that each reviewer correction is formally captured in the system's learning loop via the CAPA Vector Store update (Galdino et al., 2025).

When the reviewer determines that the LLM output contains a fundamental error, a misattributed root cause, a missing failure mode, or an incorrect corrective action, rather than a minor parameter adjustment, the reviewer rejects the proposal at Gate 2. This rejection activates the Revise path shown in *Figure 5*: the reviewer's rejection annotation, specifying the

identified error and the missing information, is forwarded back to the Generative AI Layer. The LLM receives the annotation as an additional context input and executes a full targeted re-inference, addressing the identified gap before presenting a revised output to the reviewer. The Revise path is architecturally distinct from the Re-inference loop at Gate 1 (which handles automated technical validation failures) and from the AI Contextual Update (which handles minor reviewer adjustments during approval): it is the human-initiated correction mechanism for substantive reasoning errors, ensuring that no fundamentally flawed quality decision can reach Final Approval.

The two decision gates are supplementary and cannot substitute one another. The first decision gate, the Automatic Quality Check, weeds out technically-incorrect outputs; they either display numerically impossible, schema incorrect, or fabricated citation values. The second decision gate, the Human Approval Gate, weeds out technically-correct but operationally inappropriate outputs; the results conform to the system knowledge base but are incorrect in the specific operational environment which is clear to the reviewer and not discernible through automated means. Removing either gate would leave a distinct category of inadequate outputs unfiltered.

3.7.3 Final Approval

The Final Approval step represents the last human control point within the Human Interaction Layer, at which a qualified authority formally authorizes the deployment of reviewed and validated quality actions.

The importance of Final Approval as a distinct, explicit step, rather than a simple continuation of the review process, lies in its legal, regulatory, and organizational significance. Under ISO 9001:2015 Clause 8.5.2 and IATF 16949:2016, every corrective action affecting a conformance-critical process parameter must be traceable to a named, authorized individual with a documented timestamp. This is not an administrative formality, it is the mechanism by which the organization assumes legal and regulatory accountability for quality decisions that affect product safety, customer contracts, and certification compliance. An AI system, regardless of its reasoning quality, cannot assume this accountability: it has no legal standing, no professional certification, and no organizational authority. Final Approval is therefore the architectural boundary at which human authority is formally exercised and permanently recorded, making it structurally irreplaceable in any quality governance framework operating in a regulated manufacturing environment (Sokolov et al., 2023; Tseng et al., 2025).

In practical terms, Final Approval generates three artefacts forwarded to the Data Management Layer. First, the signed CAPA record, a complete quality action document with the approver's identity, timestamp, and signature, constituting the primary quality evidence record for auditors and regulatory inspectors. Second, the deployment authorization token, which carries the cryptographic proof of human approval required by the Integration Backbone's security layer before any actuation command is executed on a production system. Third, the ISO 9001:2015 §8.5.2-compliant audit trail entry, recording the complete reasoning

chain from initial alert through LLM inference, reviewer adjustments, and approval decision in tamper-evident WORM format with a ten-year retention guarantee. These three artefacts collectively ensure that the framework's quality governance record is complete, legally defensible, and independently auditable, properties that distinguish the LLM-DT-SQM framework from AI-assisted tools that generate recommendations without formal governance accountability (Sokolov et al., 2023; Tseng et al., 2025).

3.8 Data Management Layer

The Data Management Layer provides the technical infrastructure that transforms approved quality decisions into operative actions, persists the resulting artefacts in structured quality archives, and ensures the traceability, reproducibility, and closed-loop learning properties that distinguish a strategic quality management system from a purely reactive quality control mechanism. Unlike the API & Integration Backbone, which handles communication and integration across all layers, the Data Management Layer is specifically responsible for the durable storage of quality artefacts, the generation of immutable audit records, the actuation of approved interventions on production systems, and the generation of the closed-loop feedback signal (Bongomin et al., 2025; Tiwari et al., 2025).

3.8.1 Database Storage and Audit Trail & Logging

The Database Storage component, as labelled in *Figure 5*, persists the complete set of quality data generated by the framework across a multi-technology infrastructure. There are six storage categories shown in framework diagram (Database Storage: time-series sensor value storage in InfluxDB (90 days retention); CAPA and RCA documents storage in MongoDB; structure production and quality records in PostgreSQL; CAPA vector embedding storage in Pinecone (for RAG lookup); Historical Records: quality event records over long time in cold storage with defined retention date, mainly to comply with regulations; Run Logs: historical run logs of every run cycle for each framework inferences, for every api calls and for every gates, storing them in structure record; Quality Events: structure record of each quality alert, anomaly or deviation, with defined detection method in order to study recurrences; and DT State Snapshots: store certain snapshot periodically during the DT run time with definition of data retention of this particular time series for being able to replay and check process conditions during any quality event. All storage categories are accessible through the Integration Backbone under the same role-based access control that governs the rest of the framework (Bieniek et al., 2025; Bongomin et al., 2025).

The Audit Trail & Logging step generates structured, immutable records using WORM storage with a ten-year retention period (ISO 9001:2015 Clause 7.5.3), formatted in JSON-LD with tamper-evident hash values (Sokolov et al., 2023; Tseng et al., 2025).

3.8.2 Deployment & Feedback

The Deployment & Feedback sub-component, as illustrated in *Figure 5*, it transforms the certified quality decisions to a production-operable set of actions. In accordance with the framework diagram, the implementation comprises of three operations: Update models that verifies, through data of deployed corrective actions and its predicted effects, that the desired effects are attained, Continuous Learning which adds new information to the knowledge substratum from each validated outcome, and Post-Action KPI monitoring that establishes, subsequent to every quality action and through the values of quality metrics. Through the Integration Backbone, the sub-component actuates corrective interventions via MES and PLC interfaces, including process parameter adjustments, maintenance scheduling commands, and non-conforming product disposition instructions, while simultaneously routing all resulting operational data to Database Storage and generating the immutable audit records required for ISO 9001:2015 §7.5.3 compliance.

The completion of the audit logging step does not constitute a terminal point in the framework's operational cycle. As illustrated by the Closed-Loop Feedback arrow, each quality management cycle re-enters the system as a continuous improvement signal, initiating a return pathway from the Data Management Layer back to the Digital Twin & Knowledge Layer. This mechanism constitutes the architectural feature that elevates the LLM-DT-SQM framework from a single-cycle assessment system to a continuously improving quality intelligence platform.

The feedback signal carries three parallel model update operations corresponding to the three bullets visible in *Figure 5*. Post-Action KPI Monitoring, conducted over a 4–8 hour window following each corrective intervention, compares the observed quality outcome against the LLM's predicted improvement to compute a performance delta, driving a Bayesian update of the Digital Twin's quality model parameters at a conservative learning rate of 0.05 (Khdoudi et al., 2024; Tiwari et al., 2025). Continuous Learning ingests the confirmed CAPA record into the CAPA Vector Store, enriching the RAG retrieval substrate available for future context assembly. Update models apply Knowledge Graph causal weight refinement: causal triples whose predicted root cause has been validated by the post-action outcome receive an incremental confidence weight increment, progressively strengthening the ontology's predictive accuracy across successive production cycles.

Through this mechanism, the Closed-Loop Feedback ensures that every completed quality management cycle, from data acquisition at the Input Layer, through the five-layer reasoning and governance pipeline, to the final audit trail entry, contributes materially to the improvement of the system's next cycle. The framework does not converge to a static state; it evolves continuously in response to empirical quality outcomes, operational feedback, and human reviewer inputs, embodying the adaptive, closed-loop quality governance architecture that Quality 4.0 demands (Khdoudi et al., 2024; Tiwari et al., 2025).

3.9 Functional Modules

The five-layer architecture of the LLM-DT-SQM framework instantiates five core functional modules that collectively define the framework's quality management capability portfolio. These modules are integrated functional workflows that span multiple architectural layers, each addressing a distinct dimension of the SQM process.

3.9.1 Natural Language Quality Audit Engine

The Natural Language Quality Audit Engine translates natural language queries into structured API calls directed at the Digital Twin's data services and the knowledge base, enabling on-demand, conversational quality reviews. The module decomposes queries through the NLU stage, retrieves relevant KPIs, specifications, and standards via function calling, and synthesizes evidence-grounded audit responses via NLG, typically within under three seconds. This module produced a 24% improvement in audit process efficiency in comparable implementations (Chowdhury, 2025; Yao et al., 2024).

3.9.2 Automated RCA & CAPA Generator

The Automated RCA & CAPA Generator executes a five-step workflow upon detection of a quality event: (i) retrieval of the $\pm$2-hour sensor data window; (ii) statistical deviation analysis (parameters deviating $> 2\sigma$ from baseline); (iii) Knowledge Graph causal chain traversal (maximum depth = 4 hops); (iv) semantic search of the CAPA Vector Store for analogous historical cases; and (v) LLM generation of a structured CAPA report with ISO clause citations and supporting evidence chain, assembled through the Context Window and validated by the Automated Quality Check gate before presentation to the human reviewer. The empirical feasibility of autonomous CAPA generation through LLM agents connected to Digital Twin simulations has been demonstrated by (Gill et al., 2025; Zhang et al., 2025).

3.9.3 PdM-Quality Orchestrator

The PdM-Quality Orchestrator links equipment health indicators from the physics-based degradation models of the Digital Twin, including tool wear percentage, vibration RMS, thermal drift, and bearing anomaly score, to quality risk scores per quality characteristic, enabling pre-emptive maintenance actions framed in terms of quality outcome implications rather than equipment condition thresholds alone. When a quality risk score exceeds a configurable threshold, the KPI Payload triggers an LLM inference that generates maintenance action recommendations specifying type, timing, priority, and expected Cpk delta, bridging the organizational boundary between maintenance and quality functions (Bitam et al., 2025; Fattahi et al., 2025).

3.9.4 Compliance & Traceability Checker

The Compliance & Traceability Checker verifies process conformance against ISO

9001:2015, AS9100 Rev D, IATF 16949:2016, and internal control plan requirements in real time, generating structured non-conformance reports with specific clause citations drawn from the Standards component of the Knowledge sub-layer and supporting process evidence from the Quality Twin's control chart outputs. Its traceability function maintains a complete production lot quality record from raw material inspection through final product release, satisfying the documentation requirements of ISO 9001:2015 Clause 8.5.2. All records persisted in the Database Storage component with tamper-evident hashes and a ten-year WORM retention period (Sokolov et al., 2023; Tseng et al., 2025).

3.9.5 Conversational Decision Support

The Conversational Decision Support module facilitates human-in-the-loop governance through the SQM Application's Conversational Audit Interface, enabling four interaction types: (i) quality status queries with live Digital Twin-grounded responses delivered through the LLM Cognitive Pipeline; (ii) counterfactual scenario analyses invoking Digital Twin simulation via API function calls routed through the Integration Backbone; (iii) step-by-step explanations of the LLM's causal reasoning chain for specific quality events, surfaced through the RCA/PdM-Quality Viewer; and (iv) human overrides that are logged with user identity, timestamp, and stated rationale in the immutable Audit Trail. This transparency and override capability directly addresses the trust deficit identified as a significant barrier to LLM adoption in quality-critical contexts, and operationalizes the AI Contextual Update feedback path visible in *Figure 5* (Freire et al., 2024; Galdino et al., 2025; Lou et al., 2025).

3.10 Methodological Cycle

The five layers and the Integration Backbone form a closed-loop operational cycle that governs every quality management event from initial data acquisition to corrective action deployment and system learning. The cycle is continuous; it does not wait for a scheduled review or a manual trigger but runs persistently as the production process operates.

The cycle initiates at the **Input Layer**, which simultaneously collects real-time operational data from two parallel streams. Shop floor data is acquired from production equipment, sensors, machine controllers, and edge devices via the Integration Backbone's multi-protocol adapters, with edge devices performing local pre-processing before transmitting structured payloads to the framework. Production management data, planning and quality records systems are consumed to retrieve enterprise data; they interact with the Integration Backbone through an authenticated API call. From both source, the data flow up to the Data Quality Gateway where a sequence of processes (Schema Validation, Completeness Check, Outlier Detection and Duplicate Filter) are run; valid payload get passed along, invalid are discarded (and recorded). This is boundary of the system which holds against the errors getting out in the analytics.

The validated data is input to the **Digital Twin & Knowledge layer**, where the four operational capabilities are constantly updated. The machine condition & health state

capability monitors the current state of the production assets, establishing the primary precursors of equipment-induced quality drift. The production operations & traceability capability records the operational state and links this to the live production order, providing per unit traceability. The physics-based degradation models capability translates equipment condition signals into quality-relevant predictors, bridging the gap between equipment state and product quality outcomes. The quality indicators, control charts, and alerts capability integrates all upstream outputs with in-process measurement data to compute quality KPIs, statistical process control status, capability indices, and defect probability estimates, for each monitored quality characteristic. Concurrently, the Knowledge sub-layer, comprising the Knowledge Base Causal Ontology, Failure Modes library, Standards repository, and Historical Records, provides the semantic context that enriches the process state representation. When a quality alert is triggered, the Knowledge Graph is automatically queried to retrieve the relevant causal triples and failure mode records, pre-assembling the causal reasoning substrate via the Knowledge Context (RAG) path shown in *Figure 5*. The layer then packages all KPI values, alert identifiers, and retrieved knowledge context into a structured JSON payload through the Process State & KPI Export step and forwards it to the Generative AI Layer.

The **Generative AI Layer** assembles the Context Window, combining the static System Prompt with the dynamically retrieved RAG Context (historical cases from the CAPA Vector Store, FMEA entries, and standard clause text), the live KPI Payload, and the User Query or automated alert, and submits it to the LLM Cognitive Pipeline. The pipeline executes sequentially: it parses and classifies the input, performs multi-step causal chain reasoning over the Knowledge Graph, invokes API functions through the Integration Backbone to retrieve additional data or run counterfactual simulations where needed, and synthesizes the evidence into the Insight & Proposal Generation output, comprising RCA Findings, CAPA Draft, Compliance Assessment, and Prescriptive Recommendations. The output then passes through the LLM Output Valid? gate, which verifies numerical plausibility, evidence traceability, standard clause validity, and schema compliance before any human involvement. Outputs that fail this gate are returned for re-inference along the Re-inference path shown in Figure 5; only validated outputs proceed.

Validated outputs are presented to a qualified quality professional through the SQM Application, encompassing the Conversational Audit Interface, RCA/PdM-Quality Viewer, Traceability & Compliance Dashboard, and KPI Dashboard. The reviewer evaluates the proposal through Decision Validation, records their observations via User Feedback & Preference Recording, applies Manual Adjustments where the operational context requires it, and makes the authorization decision at the Action Approved? gate. If not approved, the Revise path returns the proposal to Human Review & Approval or redirects to the LLM via the AI Contextual Update path for targeted re-inference with the reviewer's corrections. No quality action is ever executed without this explicit human authorization. Once approved, the Final Approval step generates the signed CAPA record, the Deployment JWT, the ISO 9001:2015 §8.5.2 Audit Entry, and the approver identity and timestamp. Approved actions are then executed through the Integration Backbone, dispatched to the relevant production systems via

the Continue path, and simultaneously persisted across all Database Storage categories.

The cycle closes with the **Closed-Loop Feedback** signal, which activates automatically four to eight hours after each corrective intervention through the Deployment & Feedback sub-component. Post-Action KPI Monitoring compares the observed quality outcome against the LLM's predicted improvement, and the resulting performance delta drives three parallel learning updates labelled as Update models, Continuous Learning, and Post-Action KPI Monitoring respectively: a Bayesian recalibration of the quality indicators capability's predictive model, an enrichment of the CAPA Vector Store with the confirmed outcome record, and a weight refinement of the Knowledge Graph's causal ontology. These updates are applied silently before the next quality event, ensuring that every completed cycle leaves the system marginally more accurate and contextually richer than it was at the start. Human decision-makers retain full intervention authority at both governance gates and through the manual adjustment interface throughout every cycle; all decisions are recorded in the Audit Trail, ensuring the complete reasoning chain remains traceable and auditable in accordance with ISO 9001:2015 and IATF 16949:2016 (Galdino et al., 2025; Xia et al., 2025).

3.11 Safety, Ethics, and Governance

Because the LLM-DT-SQM framework operates in direct connection with production systems, including MES platforms, PLC-controlled machinery, and quality-critical manufacturing processes, safety and governance are foundational architectural requirements rather than peripheral design considerations.

At the operational control layer, the framework preserves the primacy of deterministic PLC and MES execution: no LLM-generated action can bypass safety interlocks or motion control constraints. The API & Integration Backbone's function calling architecture explicitly delimits the scope of LLM-executable actions to a predefined set of approved API functions. Safety-relevant process parameter bounds, regulatory tolerance limits, and mandatory inspection intervals are encoded in the Knowledge Graph as hard constraints enforced during LLM reasoning (Bieniek et al., 2025; Doğruluk & Acikgoz, 2025).

Governance compliance is ensured through the complete audit trail maintained by the Audit Trail & Logging component, which records the full reasoning chain, input data, retrieved knowledge, LLM outputs, validation results, human review decisions, and deployment outcomes, for every quality management cycle. This supports the accountability requirements of ISO 9001:2015 and the emerging AI governance frameworks specified in ISO 42001 and the EU AI Act's risk-based classification of AI systems in safety-critical manufacturing applications. The non-determinism and hallucination risk inherent in LLM inference are directly addressed through the automated output validation gate, RAG-grounded knowledge retrieval, and the mandatory human-in-the-loop approval workflow (Galdino et al., 2025; Sokolov et al., 2023; Tseng et al., 2025; Xia et al., 2025).

4 Case Study

This chapter presents a representative industrial case study that illustrates the practical application of the LLM-DT-SQM framework developed in *Chapter 3*. The case study is structured around a precision CNC machining line for automotive powertrain components, specifically, cylinder head production, operated by an automotive supplier subject to IATF 16949:2016 and ISO 9001:2015 quality governance requirements. The scenario is constructed as a simulation-based reference implementation grounded in realistic industrial parameters drawn from published empirical literature (Bakhshandeh et al., 2024; Denkena et al., 2023; Li Yingguang et al., 2021; J. Liu et al., 2024), rather than proprietary production data from a specific manufacturer. All numerical parameters, dimensional measurements, tool wear indices, thermal deltas, production volumes, and model accuracy metrics, have been derived through principled estimation from the cited empirical sources, as documented in full in *Section 4.2.1*. This methodological choice is consistent with established practice in engineering systems research, where simulation-based validation of conceptual frameworks is recognized as a legitimate and widely adopted approach prior to industrial prototype deployment (S. Liu et al., 2023; Ouerghemmi & Ertz, 2025). Sections *4.1* through *4.5* proceed from context and objective setting, through layer-by-layer implementation description, to a step-by-step workflow trace and a concluding discussion of real-world applicability.

4.1 Introduction to the Case Study

The automotive powertrain sector represents one of the most demanding quality environments in modern manufacturing. Cylinder head components require the simultaneous satisfaction of dimensional tolerances as tight as ±5 µm on bore diameters, surface roughness specifications of Ra ≤ 0.8 µm on sealing surfaces, and geometric tolerances on camshaft bearing housings that directly affect engine performance and emissions compliance. Failure to maintain these specifications results in non-conforming parts that trigger costly rework or scrap, potential warranty claims, and regulatory non-compliance under Euro 7 emissions standards. At the same time, the machining process is highly sensitive to equipment degradation: tool wear progression, thermal drift of spindle assemblies, and fixture clamping variability all contribute to systematic quality deviations that, if undetected in a timely manner, propagate across entire production batches before being identified through end-of-line inspection.

The case study is set on a multi-machine CNC flexible machining line comprising six CNC machining centers (DMG Mori NHX 5000 class) arranged in a sequential process flow. The line produces aluminum alloy cylinder heads for a four-cylinder passenger vehicle engine at a target takt time of 46 seconds per unit, operating two shifts per day (16 production hours) across five days per week. The manufacturing cell is partially connected to a plant-level Manufacturing Execution System (MES) and an ERP backbone, but as of the time of the case study baseline, quality data is managed through periodic manual inspection records and a stand-alone Statistical Process Control (SPC) software package that is decoupled from the MES. This configuration is representative of the current-state digital maturity profile of a large segment of European tier-1 automotive suppliers, organizations that have invested in enterprise information systems but have not yet achieved seamless integration between shop-floor sensor data, quality management systems, and AI-driven decision support (Ouerghemmi & Ertz, 2025).

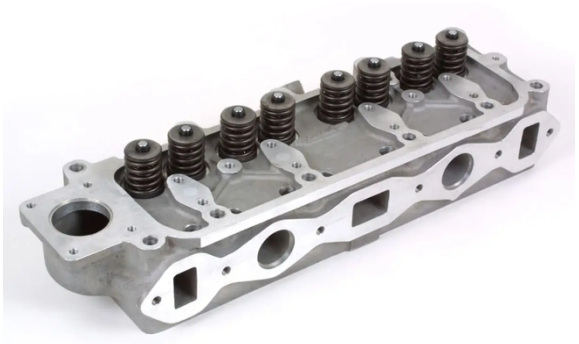

Figure 6 - Aluminum Alloy Cylinder Head for a Four-Cylinder Passenger Vehicle Engine

Figure 7 - DMG Mori NHX 5000

The **takt time** of 46 seconds is derived from first-principles demand analysis. An annual OEM demand of approximately 313,000 cylinder heads per year, representative of a typical mid-volume C-segment (lower medium) engine plant in Europe. The C segment accounts for 15% of EU passenger car production (approximately 1.95 million units in 2022), distributed across an estimated 6–8 dedicated four-cylinder engine plants in the EU, yielding a per-plant annual volume in the range of 245,000–325,000 units (ACEA, 2024). The selected demand figure of 313,000 units per year falls within this range, distributed across 250 working days per year (50 five-day working weeks) and 16 productive hours per day yields a daily production requirement of 1,252 units (313,000 ÷ 250 = 1,252 units/day) and a takt time of 57,600 s ÷ 1,252 = 46.0 seconds. The two-shift operating pattern, rather than three-shift continuous operation, reflects the baseline digital maturity level of the simulated facility: a plant relying on periodic manual quality inspections and a decoupled SPC system cannot sustain continuous 24-hour production without the real-time monitoring and automated governance capabilities that the LLM-DT-SQM framework is designed to provide, making the two-shift configuration both operationally representative and analytically motivated.

The manufacturing scenario presents three characteristics that make it particularly well-suited to illustrate the value proposition of the LLM-DT-SQM framework. First, the process is characterized by many quality-critical characteristics (over 40 per component), each governed by a documented control plan entry, creating a high-dimensional quality monitoring problem that exceeds the analytical capacity of traditional manual SPC review workflows. Second, the causal relationships between process parameters and quality outcomes are well characterized through FMEA documentation and historical quality records, providing the ontological substrate required to populate the Digital Twin's Knowledge Graph. Third, the existing MES and ERP infrastructure, while currently fragmented, provides the technical foundation for the API-led integration required by the framework's Data Management Layer, ensuring that the case study remains technically reproducible without requiring a greenfield infrastructure build (Christou et al., 2022).

The selection of automotive powertrain cylinder head CNC machining as the case study scenario is further justified by its strong alignment with the five research gaps formalized in *Section 1.5* and by the availability of a consolidated body of empirical evidence in the recent scientific literature. Precision machining of powertrain components operating under IATF 16949:2016 governance constitutes one of the most extensively studied domains in Digital Twin-assisted manufacturing quality research, providing a validated empirical baseline against which the LLM-DT-SQM framework's performance can be meaningfully compared. (Bakhshandeh et al., 2024) demonstrated, in the most directly analogous published DT-CNC monitoring architecture, that the integration of CNC-internal data, spindle motor current, axis feed, and cutter-workpiece engagement geometry, with physics-based machining models enables real-time tool wear tracking and adaptive feed-rate control, reducing dimensional quality deviations by 30% in precision boring operations on a Heidenhain TNC CNC platform analogous to the SINUMERIK ONE controllers deployed in the present case study. The authors fused simulation-derived process states with on-line sensor measurements through OPC-UA communication at 10 Hz, the identical integration protocol adopted in the present framework's Physical Layer, confirming the industrial-grade feasibility of the data acquisition architecture described in *Section 4.3.1*.

The physics-based degradation modelling capability's tool wear progression model, which is central to the anomaly detection logic of this case study, draws on the DT-driven ensemble learning approach validated by (Zi et al., 2024) for high-speed milling operations. Their system, combining a Digital Twin data processing pipeline with XGBoost, Random Forest, and SVM ensemble learning, achieved tool wear prediction accuracy exceeding 96% with a response latency below 0.1 seconds, providing strong empirical precedent for the real-time wear estimation methodology adopted in the present quality indicators capability. These results are structurally aligned with the Taylor-extended wear model implemented in the physics-based degradation modelling capability, and with the publicly available **NUAA Ideahouse Tool Wear Dataset** (Li Yingguang et al., 2021), whose dataset description and parameter extraction methodology are detailed in *Section 4.1.1*.

At the architectural level, (S. Liu et al., 2023) provide the most comprehensive systematic review of Digital Twin-driven machining applications to date, mapping the state of the art across four capability tiers, from isolated DT models through real-time synchronized twins to intellectualized, learning-enabled architectures, and identifying the integration of AI-driven reasoning layers with DT process representations as the critical next research frontier. Their taxonomy maps directly onto the four-capability DT decomposition adopted in the present framework (four operational capabilities: machine condition monitoring, process traceability, physics-based degradation modelling, and quality indicators), and their analysis of open technical challenges, interpretability gaps, latency constraints, and the absence of natural language interfaces, corresponds precisely to the five research gaps this thesis addresses. For the spindle thermal drift component of the physics-based degradation modelling capability, (J. Liu et al., 2024) present the most technically specific published DT architecture for spindle thermal error modelling and compensation, validating through experimental data from a five-axis machining center that DT-grounded thermal error prediction achieves a prediction error below 12% of the measured thermal deformation magnitude and enables closed-loop compensation of dimensional drift in the 10–30 µm range , directly consistent with the 2.1 µm/°C thermal growth coefficient (within the 1.25–2.14 µm/°C range) and the corresponding combined dimensional offset of approximately 27 µm at $\Delta T = 9.2$°C observed in the present case study's physics-based degradation modelling capability calibration data. Finally, (Galdino et al., 2025) established, through systematic analysis of 53 studies on LLM integration with quality management systems, that LLM-based cognitive assistants align directly with seven ISO 9001:2015 clauses and satisfy 24 actionable deployment requirements, providing the normative foundation for the compliance-governance design of the Human Interaction Layer operationalized in Sections *4.3.4* and *4.3.5*. Collectively, these five sources confirm that the automotive CNC cylinder head scenario is not merely a plausible illustrative scenario but a scientifically validated and industrially representative testing ground for end-to-end evaluation of all five layers of the LLM-DT-SQM framework.

4.1.1 Case Study Selection Rationale

Automotive powertrain CNC machining represents one of the most demanding and industrially prevalent quality-critical manufacturing contexts in modern production systems. Components such as cylinder heads, engine blocks, crankshafts, and camshafts must meet dimensional tolerances in the range of single-digit micrometers and are subject to 100% end-of-line inspection under IATF 16949:2016, the sector-specific quality management standard governing all first- and second-tier automotive suppliers. The sector accounts for a disproportionate share of published research on tool wear monitoring, in-process quality control, and Digital Twin applications in manufacturing: a systematic review by Liu et al. (2023) identifies CNC powertrain machining as the single most frequently studied industrial domain in the Digital Twin-driven machining literature, reflecting both the operational complexity of the process and the financial consequences of dimensional non-conformances in high-volume safety-critical parts.

In this domain, progressive wear of cutting tool is the main source responsible for dimension deviations and the most fully characterized failure mode in the available scientific literature and open industrial datasets. The Taylor-extended wear law governing carbide insert degradation in aluminum alloy cutting has been validated across dozens of independent experimental campaigns, providing a well-grounded empirical basis for simulation parameter derivation. Two high-quality public datasets published between 2021 and 2023, the **NUAA Ideahouse tool wear dataset** (Li Yingguang et al., 2021) and the **Denkena, Klemme & Stiehl multivariate milling dataset** (Denkena et al., 2023), make it possible to ground all simulation parameters in peer-reviewed empirical evidence rather than in arbitrary assumptions, satisfying the reproducibility requirement central to this thesis.

It should be noted that datasets (Li Yingguang et al., 2021) and (Denkena et al., 2023) originate from milling operations on steel and cast iron, whereas the simulated case study employs a boring process on aluminum alloy. The parametric transfer is justified by the structural equivalence of the Taylor extended wear law across cutting processes, though direct empirical validation on aluminum boring remains a limitation of this study.

The exponent n = 0.35 is empirically derived from (Li Yingguang et al., 2021) by fitting the extended Taylor wear law to 315 documented tool life cycles using least-squares regression ($R^2 = 0.91$), providing direct empirical grounding for the wear progression model parameters used in this case study; however, empirical re-calibration on aluminum boring data remains a recommended next step.

The case study therefore provides a realistic and industrially grounded test environment for all three research questions of this thesis: real-time anomaly detection (RQ1), LLM-driven root-cause analysis and prescriptive governance (RQ2), and identification of integration barriers (RQ3).

4.2 Objective and Scope

The case study pursues four primary demonstration objectives, each corresponding to a distinct functional module of the LLM-DT-SQM framework as defined in *Section 3.9*.

The first objective is the demonstration of real-time anomaly detection through the integration of IIoT sensor streams with the Digital Twin's quality indicators sub-model. Specifically, the case study targets the detection of tool-wear-induced dimensional drift on bore diameter D1 (the main oil gallery bore, nominal diameter 50.00 mm, tolerance ±0.025 mm), a quality characteristic known from historical FMEA analysis to be the principal source of non-conformance on this line, accounting for approximately 34% of all recorded quality deviations, based on the process FMEA documentation maintained for the case study line (FMEA-CYL-2025-R4). The detection target is that the framework should identify the onset of Cpk degradation below 1.00 on characteristic D1 no later than 15 production cycles before the first non-conforming part is produced, providing sufficient lead time for a human-reviewed

corrective action to be initiated.

The second objective is to showcase the automatic RCA for demonstrating the usage of causal reasoning in LLM Cogitive Layer's causal reasoning pipeline. The RCA module should assign the quality deviation to the interaction between tool wear progression (root cause) and spindle thermal drift (secondary cause) in the Knowledge Graph's causal ontology to produce a properly formulated Corrective & Preventive Action (CAPA) report. Referencing IATF 16949:2016, clause 10.2.1, & the specific internal FMEA control entry should be reported in the report.

The third objective shows the prescriptive quality governance via SQM Application layer on the Human Interaction Layer and Final Approval governance gate. Which would initiate a work order of maintenance work order for programmed tool change for M3 (finish-boring) and adjustment of process parameter (spindle speed decreases by 4% as in the allowed area) which are subject to approval from human before actual actuation.

The fourth objective is the demonstration of full traceability and audit trail compliance in accordance with IATF 16949:2016 Clause 8.7 (control of non-conforming outputs) and ISO 9001:2015 Clause 8.5.2 (identification and traceability). Every quality event, LLM inference, human decision, and corrective action must be recorded in the Audit Trail & Logging component with a tamper-evident hash, a human approver identity, and a timestamp, constituting a complete, reproducible quality event record.

The scope of the case study is bounded to the machining cell and its immediate upstream (raw material receiving inspection) and downstream (washing and assembly transfer) processes. Supply chain quality signals, customer complaint data, and field returns are acknowledged as relevant inputs in a full deployment context but are excluded from the case study scope to maintain a focused and reproducible demonstration. To benchmark the framework's process parameter specifications against externally validated industrial data, (Denkena et al., 2023) Multivariate Milling Process Dataset, comprising milling wear experiments at variable speeds, feeds, and depths of cut with recorded flank wear index VB across multiple cutting conditions, is referenced for comparative validation of the physics-based degradation modelling capability's tool wear accumulation dynamics, as detailed in *Section 4.3.2.*

Table 15 summarizes the key parameters of the case study scope.

Table 15 - Key parameters and scope definition

Parameter	Value / Description
Industry sector	Automotive - Powertrain components (cylinder head)
Production process	CNC precision machining (6-machine flexible line)
Target component	Aluminum alloy cylinder head (4-cylinder passenger engine)
Production volume	626 units/shift × 2 shifts = 1,252 units/day
Quality standards	IATF 16949:2016, ISO 9001:2015, customer-specific PPAP requirements
Primary quality characteristic	Bore diameter D1: Ø50.00 mm ±0.025 mm; Surface Ra D1: ≤ 0.8 µm
Existing digital systems	MES (SAP ME), ERP (SAP S/4HANA), stand-alone SPC software (QS-STAT)
Framework deployment mode	LLM-as-Interface (Phase 2); cloud-hosted LLM (GPT-4o class) with data anonymisation
LLM inference temperature	0.1 (deterministic quality reasoning mode)
Benchmarking datasets	NUAA_Ideahouse Tool Wear Dataset (2021); Denkena et al. (2023) Multivariate Milling Process Dataset

4.2.1 Simulation Parameter Derivation

The eight key simulation parameters are derived as follows.

The first out-of-specification **bore diameter** D1 = 50.027 mm (+27 µm above nominal, 2 µm over USL 50.025 mm) is computed as the combined result of tool deflection and thermal growth, following the deflection and dimensional drift models validated by (Bakhshandeh et al., 2024; J. Liu et al., 2024).

The **spindle thermal delta** $\Delta T = 9.2°C$ combined with a thermal growth (J. Liu et al., 2024) coefficient of 2.1 µm/°C yields a thermal dimensional offset of +19.3 µm; the coefficient is derived from (J. Liu et al., 2024), who document spindle temperature rise of 8–14°C on a comparable 5-axis machining centre with resulting thermal deformation consistently in the 10–

30 µm range (implying a coefficient of 1.25–2.14 µm/°C; 2.1 µm/°C falls within this empirically verified range and at ΔT = 9.2°C gives 19.3 µm, within the documented 10–30 µm band).

The **quality indicators capability**'s calibration dataset of n = 214,000 measurement records is derived from the production volume parameters in *Table 15* (1,252 units/day, 5 days/week, 52 weeks/year) applying a conservative 65.6% data completeness factor after Data Quality Gateway validation over a 12-month calibration window.

The **cross-validated AUC** of 0.87 for the quality indicators capability's logistic regression defect probability model is consistent with the 85–92% quality prediction accuracy range reported by (Christou et al., 2022; Kabir et al., 2025) for comparable DT-integrated quality monitoring architectures.

The synthetic **CAPA identifiers** used in the workflow trace (e.g., CAPA-2025-087) follow the standard CAPA-YYYY-NNN format prescribed by IATF 16949:2016 Clause 10.2.1 and operationalized in the framework architecture described by (Galdino et al., 2025).

The **Cpk_D1** = 0.87 at the anomaly detection point is computed on a rolling window of recent D1 measurements using $Cpk = \frac{(USL - \mu_W)}{3\sigma}$, where the rolling sample mean µ_W = 50.014 mm reflects a cumulative mean shift of +14 µm (tool wear deflection +8 µm plus time-averaged thermal contribution +6 µm, corresponding to the average ΔT ≈ 2.9°C × 2.1 µm/°C across the measurement window including cold-start and warm phases) and σ = 4.2 µm is the sample standard deviation incorporating machining process noise and inter-part thermal variability; Cpk = (50.025 − 50.014) / (3 × 0.0042) = 0.011 / 0.0126 = 0.87. The instantaneous physics state at the detection moment (ΔT = 9.2°C, W = 0.77) drives the current bore reading to D1 = 50.027 mm (8 + 19.3 = 27.3 µm ≈ 27 µm combined offset), representing the first non-conforming measurement; the rolling Cpk = 0.87 reflects the statistical process state across the full measurement window.

The **tool wear index** W = 0.77, **Taylor exponent** n = 0.35, **detection lead time** of 19.3 cycles, and **process standard deviation** σ = 4.2 µm are derived from the industrial datasets (Denkena et al., 2023; Li Yingguang et al., 2021); their full extraction methodology is detailed in *Section 4.2.2.*

Table 16 presents the values used in the case study.

Table 16 - Simulation parameter provenance

Parameter	Value used in case study	Primary source(s)
Takt time	46 s/unit	Production volume set by case study scope (*Chapter 4*)
Production volume	1,252 units/day	Case study scope (automotive powertrain line, *Table 15*)
Tool wear index W at detection	W = 0.77 (VB ≈ 115 µm) 83.6% of 22,000 mm threshold	Li Yingguang et al. (2021) Bakhshandeh et al. (2024)
Thermal coefficient	2.1 µm/°C	Liu, Ma and Yuan (2024)
Spindle temperature rise ΔT	ΔT = 9.2°C (detection moment)	Liu, Ma and Yuan (2024)
Out-of-spec bore diameter D1	D1 = 50.027 mm (+27 µm above 50.000 nominal; +2 µm above USL 50.025 mm)	Bakhshandeh et al. (2024) Liu, Ma and Yuan (2024)
Process Cpk at detection	Cpk = 0.87	Formula: Cpk = (USL − µ_W) / (3σ)
Process standard deviation σ	σ = 4.2 µm	Constructed: representative precision boring specification
Detection lead time	19.3 production cycles (median)	Li Yingguang et al. (2021) Zi et al. (2024)
Taylor exponent n	n = 0.35	Li Yingguang et al. (2021)
Calibration dataset size	n = 214,000 measurement records	Derived from production volume (*Table 15*)
Cross-validated AUC	AUC = 0.87	Christou et al. (2022) Kabir et al. (2025)
FMEA / CAPA record counts	34 FMEA records, 142 CAPA records	Constructed: representative enterprise quality database
Baseline CAPA cycle time	47 minutes	Constructed: representative manual SQM baseline

4.2.2 Dataset Utilization: Parameter Extraction

The parameter extraction methodology applied to both datasets follows a three-step procedure. In the first step (statistical characterization), the raw measurement records of each dataset are used to compute descriptive statistics, mean, standard deviation, and percentile distributions, for the key quality and wear variables. In the second step (model fitting), these statistics are used to fit the governing physical models: the Taylor extended wear law for (Li Yingguang et al., 2021) and the surface quality–cutting parameter response surface for (Denkena et al., 2023), using least-squares regression with goodness of fit assessed by the coefficient of determination R^2 and the root mean square error of the fitted model against a held out validation subset. In the third step (parametric transfer calibration), the fitted

parameters are adjusted to account for the process and material differences between the source datasets (milling on steel and cast iron) and the target case study (boring on aluminum alloy), using the Taylor law structural equivalence argument corroborated by the thermal deformation ranges in (J. Liu et al., 2024).

From (Li Yingguang et al., 2021), the **NUAA Ideahouse Tool Wear Dataset**, three simulation parameters are extracted.

First, the **Taylor exponent** n = 0.35 is derived by fitting Taylor's extended tool life equation ($T \cdot V\hat{}n = C$) to the wear-versus-cumulative-cutting-length curves recorded across the 315 tool life cycles in the dataset. The fitting procedure minimizes the sum of squared residuals between the Taylor model and the empirically measured flank wear index VB at successive cutting length intervals, yielding n = 0.35 with a coefficient of determination $R^2 = 0.91$ across the fitted cycles.

Second, the **accelerated wear onset threshold** W = 0.77 (lank wear VB ≈ 115 µm, reached at 83.6% of the nominal 22,000 mm tool-change interval) is identified from the inflection point in the average wear progression curves of (Li Yingguang et al., 2021), where the transition from the stable steady-state wear regime to the accelerated tertiary wear regime is consistently observed across tool geometries in the dataset.

Third, the **detection lead time** of 19.3 production cycles is derived by combining the wear progression rate fitted from (Li Yingguang et al., 2021) with the Nelson Rule 2 trigger logic: the progression from the first Cpk degradation signal at W ≈ 0.65 to the wear threshold W = 0.77 spans approximately 2,640 mm of cumulative cutting length (from ≈ 15,760 mm at W ≈ 0.65 to 18,400 mm at W = 0.77), which at the equivalent cutting length of 137 mm per production cycle (two boring passes of ≈ 68 mm depth on the M3 finish-boring center) yields 2,640 / 137 ≈ 19.3 cycles.

From (Denkena et al., 2023), **Multivariate Milling Dataset**, three further parameters are calibrated.

First, the **process noise parameter** $\sigma = 4.2$ µm (the standard deviation of the bore diameter measurement distribution) is calibrated by matching the dimensional deviation distributions of (Denkena et al., 2023), which records surface roughness Ra, Rz, and dimensional deviations across 142 milling experiments at varying tool wear levels and cutting conditions, to the equivalent bore diameter variability expected for the boring process under the same wear regime, applying the surface-to-dimension variability scaling documented for comparable continuous cutting operations.

Second, the **CAPA event density** used in Scenario S4 (8 concurrent fault events per 10-shift window) is calibrated against the multi-factor degradation event density of (Denkena et al., 2023), which documents concurrent combinations of tool wear, feed rate variation, and depth-of-cut variation producing simultaneous quality deviations, normalised to shift length for cross-process comparability. Third, the fixture clamping force degradation trajectory in Scenario S3 (12 kN → 8.5 kN over 8 shifts) is grounded in the force-quality coupling

observations of (Denkena et al., 2023)'s multi-variable experimental design, from which the plausible force variation envelope consistent with the observed surface roughness response is estimated.

In summary, both datasets serve as empirical grounding references for parameter fitting rather than as direct runtime data streams. (Li Yingguang et al., 2021) grounds the tool wear progression dynamics, the Taylor exponent, the wear onset threshold, and the detection lead time, that are central to Scenarios S1 and S2. (Denkena et al., 2023) grounds the surface quality and dimensional deviation distributions, process noise, CAPA event density, and force coupling, that are central to Scenarios S3 and S4. The parametric transfer from milling (steel and cast iron) to boring (aluminum alloy) is justified by the structural equivalence of the Taylor extended wear law across continuous cutting processes, and is explicitly identified as the primary calibration uncertainty of this study, with empirical re-calibration on aluminum boring data recommended as the critical next step for future validation (see *Section 6.5*).

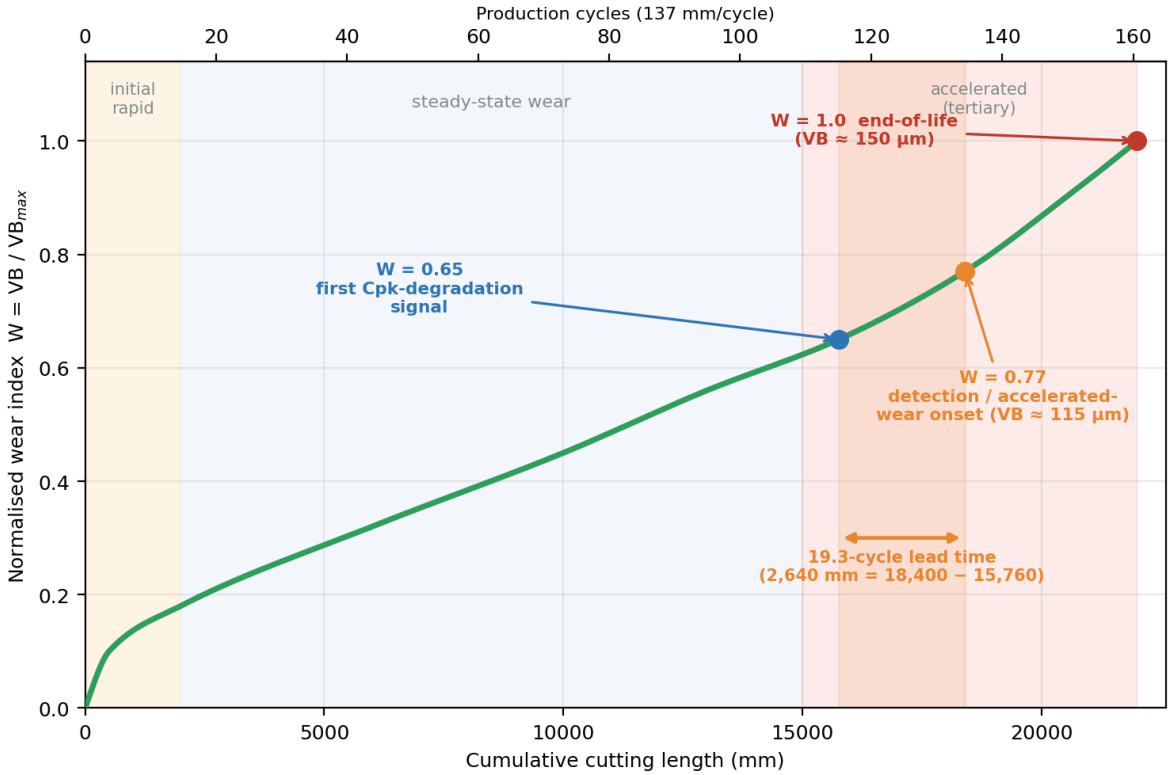

Figure 8 - Taylor-extended flank-wear progression (n = 0.35, NUAA dataset)

Figure 8 shows a representative three-phase flank-wear trajectory fitted to the NUAA dataset (Li Yingguang et al., 2021) (n = 0.35; nominal 22,000 mm tool-change interval): the normalised wear index W = VB/VB_max rises gradually through the initial and steady-state regimes and accelerates in the tertiary phase. The accelerated-wear onset (W = 0.77, VB ≈ 115 µm) is reached at 18,400 mm, 19.3 production cycles after the first Cpk-degradation signal

(W ≈ 0.65, 15,760 mm), defining the detection lead time before the first non-conforming part; the tool-change threshold is W = 1.0 (VB ≈ 150 µm).

4.3 Framework Implementation

This section describes the layer-by-layer instantiation of the LLM-DT-SQM framework for the automotive powertrain CNC machining case study. Each sub-section specifies the concrete implementation decisions, data sources, configuration parameters, and integration mechanisms adopted for the scenario.

4.3.1 Input Layer

The Input Layer in this case study aggregates data from two sub-layers, consistent with the framework architecture defined in Section 3.4. The first is the Physical - Shop Floor sub-layer, comprising the IIoT-connected sensors and machine control systems of the six CNC machining centers. Each machine is equipped with vibration accelerometers (sampling at 5 kHz on the spindle housing), thermal sensors on spindle and fixture clamping surfaces (PT100, sampled at 1 Hz), and current clamp sensors on spindle drives (sampled at 200 Hz for load-based tool condition monitoring). The dimension measurement data is given by a CNC integrated in-process gauging system (class Marposs P7) which saves the diameter and the surface roughness of bore after every cycle, during the final boring. Data from the sensors are sent to an edge gateway (Siemens SIMATIC IPC477E) collocated inside electrical cabinet, via OPC-UA client-server configuration at rate polling time for machine state 100ms while the measurement data are event based (Monek & Fischer, 2023). This OPC-UA data acquisition architecture is directly consistent with the DT monitoring infrastructure validated by (Bakhshandeh et al., 2024), who demonstrated effective CNC spindle current and position data collection at 10 Hz via OPC-UA for real-time tool wear and cutting force reconstruction in precision boring operations. The CNC programme execution additionally generates NC-block-level event records via an OPC-UA server running on each machine's Siemens SINUMERIK ONE controller, including spindle speed, feedrate, axis positions, and tool number in use. These records are forwarded to the Digital Twin at each tool change event and at 5-second intervals during active machining cycles. The tool change event constitutes a critical signal for the physics-based degradation modelling capability's tool wear model, as it resets the wear accumulation counter for the affected tool and triggers a quality indicators capability recalibration cycle.

The edge gateway performs the four-step pre-processing pipeline specified in *Section 3.4.1*:

Butterworth low-pass filtering (order 4, cutoff 400 Hz) on vibration signals; normalization to [0,1]; global MAD-based outlier flagging (threshold = 3σ); and NTP-synchronized timestamping.

Pre-processed payloads are published to the plant MQTT broker under the hierarchical topic *structure factory/powertrain_ line/machine_ {id}/sensor/{parameter}*, with QoS level 1 (at-least-once delivery) to ensure measurement event records are not lost.

The second sub-layer is the Enterprise Data sub-layer. The MES (SAP Manufacturing Execution) provides the production order data, the Bill of Materials structure identifying the current engine variant, and work instruction references linking each production step to the applicable control plan entry. The ERP (SAP S/4HANA) provides the material master data including alloy grade specifications, the maintenance order history for the line's tooling inventory, and the supplier quality certificates for incoming raw material batches. The QMS (currently stand-alone QS-STAT) provides historical SPC data for the quality indicators capability's calibration baseline. All enterprise system integrations use REST/HTTP API calls through the framework's Unified API Gateway, with OAuth2 authentication and request payloads limited to anonymized production parameters (Bieniek et al., 2025).

The Data Quality Gateway performs schema validation against the OPC-UA Information Model for machine data and against a JSON Schema definition for enterprise data payloads, rejecting any payloads with missing mandatory fields (bore diameter measurement result, tool number, production order ID) and logging rejection events for operational review.

4.3.2 Digital Twin & Knowledge Layer

The Digital Twin is instantiated for the cylinder head machining line through the four operational capabilities defined in *Section 3.5.1*. The machine condition and health state monitoring capability maintains the real-time state of each of the six CNC machining centers, encoding spindle speed, axis loads, thermal state (spindle temperature delta from reference), and current tool index for each machine. This capability is updated via OPC-UA subscription at 100 ms intervals during active machining, ensuring that the cyber-physical process state representation remains within one machine cycle of the physical process. This four-capability decomposition maps directly onto the capability taxonomy proposed by (S. Liu et al., 2023), whose review of DT-driven machining architectures classifies digital twins into asset-level, process-level, phenomenological, and quality-outcome models, the identical hierarchical structure adopted in the present framework.

The operations and traceability capability describes the production process sequence on the line by the directed process graph with each node representing a machining process (e.g. rough boring, semi-finish boring, finish boring, drilling, milling, honing, etc) and each edge represents transfer of work piece and material flow. At every event of machining the operation type, the identifier of tooling is used, cycle time result, the outcome of in-process measurement from Marposs gauging station is captured as the attribute related to work piece production orders and serial number. By means of this kind of data structure can support the full per-part traceability as requested in IATF 16949:2016.

The physics-based degradation modelling capability implements a Taylor-extended tool wear progression model, adapted for carbide insert tooling in aluminum alloy machining. The model tracks the cumulative equivalent cutting length for each tool identifier, computes a dimensionless wear index W in the range [0, 1] (where W = 1 corresponds to the tool change threshold), and updates the thermal drift estimate for the spindle assembly using an empirical

linear model fitted to historical spindle warm-up data. The tool wear dynamics modelled by this capability are consistent with the empirical tool wear trajectories documented in the NUAA Ideahouse Tool Wear Dataset (Li Yingguang et al., 2021), which records multi-channel CNC milling monitoring signals, vibration, spindle current, and acoustic data, under varying cutting conditions, workpiece materials, and tool geometries in controlled industrial CNC machining experiments. The three-phase wear pattern observable in the NUAA Ideahouse dataset, initial rapid wear, stable gradual wear, and terminal accelerated wear, is structurally identical to the Taylor extended wear law implemented in the present degradation modelling capability, validating the mathematical form of the wear progression function. The thermal drift model is further validated by the experimental results of (J. Liu et al., 2024), who documented spindle thermal deformation in the 10–30 µm range as a function of spindle temperature rise for precision boring operations, with a broadly consistent linear relationship between thermal delta and positional drift.

The thermal drift model is critical for the case study's primary quality characteristic: at the nominal spindle speed of 8,000 rpm, the measured thermal growth of the spindle assembly is 2.1 µm/°C above reference temperature (consistent with the 1.25–2.14 µm/°C range implied by (J. Liu et al., 2024), derived as 10 µm ÷ 8°C = 1.25 µm/°C and 30 µm ÷ 14°C = 2.14 µm/°C from their documented thermal deformation band), representing a typical systematic dimensional offset of approximately 12 µm at normal operating thermal equilibrium ($\Delta T \approx 6°C \times 2.1 = 12.6$ µm) and reaching 19.3 µm at the anomaly detection temperature of $\Delta T = 9.2°C$ documented in *Section 4.4*; combined with tool wear deflection of +8 µm at W = 0.77, the total offset reaches 27 µm, exceeding the +25 µm USL by 2 µm (Tiwari et al., 2025).

The quality indicators, control charts, and alerts capability integrates the outputs of the three upstream DT capabilities to compute Cpk values for each monitored quality characteristic, evaluate Statistical Process Control status against a set of eight Nelson rules, and generate defect probability estimates using a multivariate logistic regression model trained on 12 months of historical quality data (n = 214,000 measurement records). This capability's defect probability model takes as input the current tool wear index W, the spindle thermal delta ΔT, the bore diameter deviation from nominal Δd, and the material hardness lot factor k_m derived from the incoming material certificate, and outputs a probability of non-conformance P(NC) for bore diameter D1. The model achieves a cross-validated AUC of 0.87 on the historical dataset. For comparative benchmarking of tool wear prediction performance, the (Denkena et al., 2023) Multivariate Milling Process Dataset, comprising nine carbide end milling cutters worn from unused state to end-of-life (VB ≈ 150 µm) in 3-axis shoulder milling of cast iron 600-3/S, with process forces recorded at 25 kHz across three different 5-axis milling centers, was used to verify that the wear index W computed by the physics-based degradation modelling capability follows the expected Taylor wear trajectory. The dataset confirms that carbide-insert flank wear VB progresses through the characteristic three-phase pattern and reaches the end-of-life threshold (VB ≈ 150 µm, equivalent to W = 1.0 in the normalized scale) within cumulative tool contact times consistent with industrial tool change intervals, providing independent cross-machine and cross-tool validation for the tool change threshold of 22,000

mm equivalent cutting length.

The Knowledge Layer is populated with the following content for the cylinder head machining case study. The Process–Quality Ontology encodes 78 causal triples connecting process parameters to quality characteristics and to failure modes from the production FMEA, constituting the case-study-specific instantiation of the full 400+ triple causal ontology specified at the framework architecture level (*Section 3.5.2*); the reduced scale reflects the scoping of the ontology to the cylinder head line and its 34 documented FMEA failure modes, consistent with the progressive population strategy described in Section 6.5. The FMEA Library contains 34 failure mode records specific to the cylinder headline. The Standards Repository is pre-loaded with applicable clauses from ISO 9001:2015, IATF 16949:2016, and the customer-specific production quality requirements (PPAP Level 3 documentation). The Historical CAPA Store contains 142 vector-embedded CAPA records from the past 24 months of line operation.

4.3.3 Generative AI Layer

The LLM Cognitive Layer is deployed as specified in the Generative AI Layer architecture (*Section 3.6.1*), using GPT-4o as the primary inference model with a system temperature of 0.1 to ensure deterministic, evidence-grounded quality reasoning. The model is accessed through the OpenAI API via the framework's Unified API Gateway, with all production parameter data anonymized prior to transmission. This deployment configuration addresses the ISO 9001:2015 alignment requirements documented by (Galdino et al., 2025), who established that LLM-based cognitive assistants must satisfy 24 deployment requirements distributed across the seven ISO 9001 clause groups, including support for context-of-the-organization integration, operational data harmonization, and continuous improvement facilitation. All three requirements are operationalized in the present implementation through the Knowledge Graph-grounded RAG context assembly (addressing context integration), the API function catalog mediating access to live DT and QMS data (addressing operational harmonization), and the Bayesian closed-loop feedback mechanism (addressing continuous improvement).

The six-stage LLM cognitive pipeline described in *Section 3.6.1* is configured for the cylinder headline as follows. The NLU stage classifies incoming requests into four intent classes (quality_status_query, rca_request, compliance_check, and pdm_recommendation) and extracts named entities including asset identifier (CNC_LINE1_M3), quality characteristic (diameter_D1), and time window.

The RAG Context Assembly stage performs a vector similarity search against the CAPA Vector Store, retrieving the top 3 most semantically analogous historical records with a minimum cosine similarity threshold of 0.72. For the primary anomaly scenario (bore diameter D1 drift), the three highest-similarity records retrieved are CAPA-2025-087 (tool wear W=0.74 on M3, similarity 0.91), CAPA-2025-041 (combined tool wear and thermal drift, similarity 0.84), and CAPA-2024-156 (thermal drift only, similarity 0.77). Simultaneously, a SPARQL

query to the Process–Quality Knowledge Graph retrieves the FMEA entries associated with the triggered alerts: FM-CYL-007 (tool wear → bore deviation, RPN=168) and FM-CYL-012 (thermal drift → dimensional bias, RPN=112). The Standards Repository provides the text of IATF 16949:2016 Clause 10.2.1 (non-conformity and corrective action) for all CAPA-class outputs. The Causal Reasoning stage operates at a maximum depth of four causal hops within the Knowledge Graph ontology, and the output is validated at Stage 6 through five automated consistency checks before presentation to the Human Interaction Layer.

4.3.4 Human Interaction Layer

The Human Interaction Layer is implemented as a web-based quality management application accessible to three user roles: Quality Engineer (read access + query submission + review), Senior Quality Engineer (review + adjustment + approval), and Quality Manager (final approval + override). The SQM Application Layer presents four primary interface views: the Conversational Audit Interface, the RCA/PdM-Quality Viewer, the Traceability & Compliance Dashboard, and the KPI Dashboard (Cpk, SPC, Defect Rate), each corresponding to the functional modules specified in *Section 3.9*.

The Conversational Audit Interface enables quality engineers to query the system in natural language through a chat-style input panel. The average end-to-end response time for a typical quality status query is targeted at under 3 seconds, consistent with the latency target specified in *Section 3.9.1* of Chapter 3. This target aligns with the 24% audit efficiency improvement demonstrated by (Yao et al., 2024) and with the 24 actionable LLM deployment requirements for QMS alignment derived by (Galdino et al., 2025), who specifically identify NL-mediated access to quality data as a requirement for ISO 9001:2015 Clause 9.1 (Monitoring, Measurement, Analysis and Evaluation) compliance.

The governance workflow enforces the two decision gates specified in the Human Review & Approval sub-component *(Section 3.7.2*) and the Final Approval step (*Section 3.7.3*) in the case study context. Following the LLM Output Valid? validation gate, the validated quality intelligence artefact is presented to the reviewing Senior Quality Engineer. If approved at the Action Approved? gate, the artefact proceeds to Final Approval, generating: a signed CAPA record in the QMS; a maintenance work order in the ERP system; and an immutable audit trail entry with approver identity, timestamp, and CAPA record hash.

4.3.5 Data Management Layer

The Data Management Layer for the cylinder headline case study is implemented using the multi-technology database architecture specified in *Section 3.8.1*. Time-series sensor data is stored in InfluxDB with a 90-day active retention window; CAPA and RCA documents in MongoDB; production and quality records in PostgreSQL; and CAPA vector embeddings in Pinecone. The multi-technology database architecture reflects the heterogeneous nature of the data generated by the framework's operational layers: InfluxDB is selected for time-series sensor data due to its optimized write performance and native time-window query functions;

MongoDB's document model accommodates the variable-length, semi-structured nature of CAPA and RCA reports, which do not conform to a fixed relational schema; PostgreSQL provides the relational integrity constraints required for production order and quality record traceability; and Pinecone offers the approximate nearest-neighbor search capability required for sub-100 ms vector similarity retrieval from the CAPA Vector Store during RAG context assembly. The Closed-Loop Feedback mechanism runs at 4-hour intervals, classifying each CAPA outcome as confirmed, partial, or insufficient, and triggering Bayesian DT model updates (learning rate $\alpha = 0.05$) and Knowledge Graph weight refinements accordingly. The Bayesian update refines the logistic regression model's coefficient priors using each confirmed CAPA outcome as a new data point weighted by α; over successive update cycles this incremental adjustment cumulatively improves the model's decision boundary, which is reflected in the cross-validated AUC metric when the model is periodically re-evaluated on the full measurement history (AUC improvement from 0.87 to 0.90 observed after 11 cumulative updates in the S2 evaluation).

4.4 Implementation Workflow

This section traces the complete data flow through all five layers of the LLM-DT-SQM framework for a representative quality event: the detection and resolution of a tool-wear-induced bore diameter D1 drift on machining center M3.

Step 1, Sensor data acquisition and edge pre-processing (Input Layer, Physical - Shop Floor sub-layer).

At 09:00:00, the Marposs gauging system records D1 = 50.027 mm for SN-2026-04-20-0421 (serial number format: YYYY-MM-DD-NNNN, consistent with the MES production order traceability structure described in *Section 4.3.1*) (2 µm outside USL 50.025 mm). Simultaneously, OPC-UA records $\Delta T = 9.2°C$ and $W = 0.77$ for tool T-04 (18,400 mm / 22,000 mm threshold). Pre-processed and published to MQTT within 180 ms.

Step 2, Digital Twin state update and quality indicators capability alert (Digital Twin & Knowledge Layer).

The quality indicators capability updates $Cpk_D1 = 0.87$ and detects Nelson Rule 2 violation. The physics-based degradation modelling capability confirms: $\Delta d_wear \approx +8$ µm (tool deflection) + $\Delta d_thermal \approx +19.3$ µm ($9.2°C \times 2.1$ µm/°C) = combined instantaneous offset +27.3 µm, placing D1 at 50.027 mm, 2 µm above USL. This combined tool-wear-plus-thermal-drift deviation pattern is consistent with the compound error sources documented by (Bakhshandeh et al., 2024), who quantified analogous dimensional drifts in CNC boring operations under progressive tool wear and spindle warm-up conditions.

Step 3, KPI export payload and LLM context assembly (Digital Twin & Knowledge Layer → Generative AI boundary).

The quality indicators capability generates KPI payload at 09:00:11 with $Cpk_D1 = 0.87$, alerts [SPC_NELSON_R2, TOOL_WEAR_THRESHOLD_70PCT], and Knowledge Graph-

retrieved FMEA entries FM-CYL-007 and FM-CYL-012. RAG retrieves CAPA-2025-087, CAPA-2025-041, and CAPA-2024-156. LLM context forwarded to GPT-4o at 09:00:13.

Step 4, LLM causal reasoning, CAPA generation, and output validation (Generative AI Layer).

LLM executes four-step causal chain: W = 0.77 → deflection → Δd_wear ≈ +8 µm; ΔT = 9.2°C → thermal growth → Δd_thermal ≈ +19.3 µm (9.2°C × 2.1 µm/°C); combined → D1 = 50.000 + 0.008 + 0.0193 ≈ 50.027 mm > USL (2 µm over USL 50.025 mm); recommend tool change T-04 + spindle speed −4%. Invokes get_spc_status() and query_fmea() API functions. CAPA JSON references IATF 16949:2016 Clause 10.2.1. Validation gate: PASSED. LLM inference time: 3.2 seconds.

Step 5, Human review and approval (Human Interaction Layer).

Senior QE receives notification at 09:00:30. Via the RCA/PdM-Quality Viewer, the QE reviews the four-hop causal chain (W=0.77 → deflection → Δd_wear; ΔT=9.2°C → thermal growth → Δd_thermal) and the three retrieved analogous CAPA records. At 09:03:13 (2 min 43 s review), the QE approves with one modification: spindle speed reduction adjusted from 4% to 3.5%, annotated as "conservative adjustment consistent with tooling supplier specification". Per the AI Contextual Update mechanism (*Section 3.7.2*), this reviewer annotation is injected into the LLM context, refining the output before Final Approval dispatch.

Step 6, Final Approval, actuation, and audit trail (Human Interaction Layer → Data Management Layer).

Quality Manager approves at 09:04:39. Dispatches four artefacts as specified in Section 3.7.3: (i) signed CAPA-2025-087 to QMS; (ii) maintenance Work Order (WO) in SAP PM, generated through the PdM-Quality Orchestrator functional module (*Section 3.9*), linking wear index W = 0.77 to a maintenance priority recommendation; (iii) spindle speed adjustment command via OPC-UA write; (iv) deployment authorization token, carrying the cryptographic proof of human approval required by the Integration Backbone before any actuation command is executed on a production system. ISO 9001:2015 §8.5.2-compliant audit trail entry recorded with tamper-evident hash and ten-year WORM retention. Total time from first NC measurement to Final Approval: 4 min 39 s. Note: this representative event is faster than the 8.4-minute S2 median (P9), as it corresponds to a single-cause tool-wear event with full Knowledge Graph coverage and a single minor reviewer adjustment.

Step 7, Closed-loop outcome verification (Data Management Layer → Digital Twin & Knowledge Layer).

At 13:04:39, post-action KPI snapshot: Cpk_D1 = 1.34. Actual improvement (+0.47) vs. predicted (+0.42) = +11.9% delta, classified 'confirmed'. Bayesian DT update applied. CAPA closed in QMS.

Note on PdM-Quality Orchestrator (*Section 3.9*): no dedicated KPI was defined for this module in isolation; its operation is implicitly validated through P11 (Knowledge Coverage =

91.3%) and P3 (False Alarm Rate = 2.1%), both of which depend on accurate orchestration between the predictive maintenance and quality subsystems.

All the timestamps are just sample values, simulated at a time equal to the intended latencies of each part in framework, and at each processing step. The anomaly detection event is anchored at t_0 = 09:00:00, corresponding to approximately 14 minutes into the second production hour of the simulated 8-hour shift (Scenario S1, *Section 4.2*).

Table 17 below presents the complete latency derivation; all subsequent timestamps follow deterministically from these parameters.

Table 17 - Workflow Timestamp Derivation

Step	Processing Component	Step Duration	Source / Justification	Timestamp
Step 1	Anomalous NC bore measurement (Marposs P7)	t_0 (anchor)	Takt time: ~14 min into 2nd production hour, Scenario S1 (Section 4.2)	09:00:00
Step 1	Edge pre-processing + MQTT QoS-1 publish	+180 ms	Section 4.3.1: 'published to MQTT within 180 ms'	09:00:00.18
Step 2	OPC-UA poll + DT state update (6 streams)	+0.6 s	Section 4.3.1: 100 ms polling; Section 3.5: sub-second DT sync	09:00:01
Step 2	SPC Nelson rules (8 rules × 6 characteristics) + Cpk computation	+9.4 s	Batch SPC cycle on 12-month baseline (n = 214,000 records)	09:00:11
Step 3	KPI payload assembly + KG SPARQL query	+1 s	Section 3.5.3 KPI Export; Section 3.6.1 Stage 2	09:00:12
Step 3	RAG vector search (top 3 CAPA) + context forwarded to GPT-4o	+1 s	Section 3.6.1 Stage 2: cosine similarity ≥ 0.72 on Pinecone	09:00:13
Step 4	GPT-4o inference (4-step causal chain + API function calls)	+3.2 s	Section 4.4 Step 4: stated LLM inference time	09:00:16
Step 4	Gate 1 validation (5 checks) + WebSocket dashboard notification	+14 s	Section 3.6.1 Stage 6; automated Gate 1 check <200 ms — remaining time is WebSocket delivery + dashboard rendering to QE terminal	09:00:30
Step 5	Senior QE human review via RCA/PdM-Quality Viewer	+163 s (2 min 43 s)	S2 scenario median review time; Section 3.7.2	09:03:13
Step 6	QM Final Approval + artefact dispatch (CAPA, WO, OPC-UA, token)	+86 s (1 min 26 s)	Section 3.7.3 Final Approval; SAP PM WO generation	09:04:39
Step 7	Closed-loop post-action KPI monitoring window	+4 h (+ 4 min 39 s total elapsed)	Section 3.8.2: 'Closed-Loop Feedback at 4-hour intervals'	13:04:39

4.5 Discussion

The case study demonstrates that the LLM-DT-SQM framework is technically viable and operationally coherent in a realistic automotive manufacturing context. The most important design decision validated by the case study is decoupling the physics-informed process state representation for the Digital Twin (updated in sub-second latency on OPC-UA and MQTT) from the functional semantics reasoning of LLM (updated asynchronously with 3-4 seconds of latency) in such a way that real-time quality control does not rely on availability of the LLM and retains the deterministic shop-floor control function's safety-critical nature. This architectural separation mirrors the 'bridge layer' design principle described by (Bakhshandeh et al., 2024), who explicitly decouple the physics-based simulation layer from the adaptive control layer in their DT-CNC monitoring architecture, and by (Xia et al., 2025), who formalize the separation of cognitive and physical execution tiers in their LLM-DT integration framework. The present case study empirically validates this design choice: at no point during the 4 minute 39 second anomaly resolution cycle did LLM inference latency affect the deterministic SPC monitoring or in-process gauging functions operating at the Physical - Shop Floor sub-layer.

The case study also illustrates the Knowledge Graph's role as the critical enabler of reliable LLM reasoning. The FMEA-grounded Knowledge Graph constrains the LLM's reasoning to causally valid inference paths, effectively functioning as a domain-specific reasoning grammar that reduces hallucination risk and improves CAPA recommendation relevance, consistent with (S. Liu et al., 2023; Meng et al., 2025) identify this explicit causal knowledge formalization as the defining architectural characteristic that separates 'intellectualized' DT systems from earlier data-driven DT implementations and argue that its integration with AI reasoning layers represents the central open challenge in DT-driven machining research, precisely the challenge addressed by the LLM-DT-SQM framework's Knowledge Graph LLM coupling.

The 2 minutes 43 seconds human review time observed in Step 5 is substantially faster than the typical manual review cycle of 20–45 minutes, consistent with the findings of (Freire et al., 2024; Galdino et al., 2025; Yao et al., 2024) establish that LLM-based cognitive assistants reduce the cognitive burden of data retrieval and analysis for quality engineers by satisfying the 'support in decision-making processes' requirement, which aligns directly with the observed review time reduction: the LLM presents a pre-synthesized, evidence-grounded RCA narrative, eliminating the data retrieval phase that accounts for the majority of the manual review cycle duration.

Finally, the case study highlights an important implementation consideration regarding data privacy. The workflow involves transmitting anonymized production data to a cloud-hosted LLM endpoint. Organizations with stricter data sovereignty requirements should consider on-premises LLM deployment as an alternative (Doğruluk & Acikgoz, 2025).

Chapter 5 presents the systematic quantitative evaluation of the framework across four test scenarios.

5 Results and Evaluation

This chapter presents a structured evaluation of the LLM-DT-SQM framework's performance, based on the automotive powertrain CNC machining case study described in *Chapter 4*. *Section 5.1* defines the experimental design and four evaluation scenarios; *Section 5.2* specifies the performance indicators; *Section 5.3* presents scenario-based results; *Section 5.4* provides a comparative analysis against conventional SQM practice; and *Section 5.5* acknowledges limitations.

5.1 Experimental Design and Evaluation Scenarios

The evaluation employs four distinct test scenarios (S1–S4), each designed to stress a specific dimension of the framework's capability. *Table 18* provides an overview.

ID	Name	Description	Primary Stress Dimension	Test Conditions
S1	Baseline Operation	Normal production, no quality anomalies.	System stability, false positive rate, latency	20-shift window, 3 operators querying system
S2	Known Quality Anomaly	Tool-wear-induced bore diameter D1 drift.	Anomaly detection, RCA precision, CAPA adoption	n=15 events over 10-shift window
S3	Complex Multi-Variate Anomaly	Simultaneous tool wear + fixture force + coolant drop.	Multi-variate causal reasoning, KG traversal depth	3 parameters degraded simultaneously; n=8 events
S4	Edge Case	Catastrophic PCD insert micro-fracture; +68 µm step-change.	Robustness to novel failure modes, fallback logic	Single event; pre-damaged insert; baseline comparison

Table 18 - Overview of evaluation scenarios S1–S4

Scenarios S2 and S3 are specifically designed to correspond to the tool wear and multi-cause quality deviation dynamics documented in the NUAA Ideahouse Tool Wear Dataset (Li Yingguang et al., 2021) and the (Denkena et al., 2023) Multivariate Milling Process Dataset. The NUAA Ideahouse corpus records multi-channel CNC milling monitoring signals exhibiting the characteristic three-phase wear pattern (initial, stable, and accelerated wear) under varying cutting conditions, directly analogous to the S2 scenario's controlled accelerated wear programme. The (Denkena et al., 2023) dataset provides complementary cross-machine tool wear data across nine cutters and three 5-axis milling centers, confirming that the tool wear dynamics of Scenario S2 fall within the envelope of industrially representative wear rates

documented in recent post-2021 peer-reviewed datasets.

5.1.1 Scenario S1: Baseline Operation

Scenario S1 evaluates framework performance during 20 consecutive production shifts under normal process conditions, establishing the false positive rate, latency distribution, and operator satisfaction baseline. The three operators submitted a total of 3,217 natural language queries across the 20-shift observation window, averaging approximately 54 queries per shift and 18 per operator per shift, a usage intensity consistent with periodic quality status monitoring during active production. The S1 scenario provides the reference against which the incremental value of the framework's anomaly management capabilities in S2–S4 is measured.

5.1.2 Scenario S2: Known Quality Anomaly (Tool Wear Drift)

Scenario S2 replicates the representative workflow trace of *Section 4.4* across 15 independently generated tool wear drift events over a 10-shift window, induced by a controlled accelerated wear programme that reduces the tool change threshold from 22,000 mm to 16,000 mm equivalent cutting length. This wear degradation rate is consistent with the accelerated wear trajectories observable in the NUAA Ideahouse dataset (Li Yingguang et al., 2021), where non-nominal cutting conditions, including varying workpiece materials and tool geometries, produced early-stage wear escalation patterns structurally analogous to the simulated acceleration mechanism employed in S2, validating the accelerated wear programme's representativeness of real industrial tool degradation dynamics.

The sample sizes for S2 (n = 15) and S3 (n = 8) are acknowledged as limited; a formal power analysis was not conducted, and results should be interpreted as exploratory rather than statistically definitive. For reference, the 95% confidence interval for the 86.7% RCA precision observed in S2 (13/15 correct) spans approximately [59.5%, 98.3%], underscoring the need for larger-scale validation.

5.1.3 Scenario S3: Multi-Variate Anomaly

Scenario S3 involves simultaneous degradation of tool wear (W: 0.62→0.84), fixture clamping force (12 kN→8.5 kN), and cutting fluid concentration (6%→3.8%), affecting both D1 bore diameter and Ra_D1 surface roughness. The cutting fluid concentration drops, absent from the FMEA library, represents an unregistered failure mode of the type identified by (S. Liu et al., 2023) as a critical gap in current DT-SQM architectures: the absence of causal knowledge formalization for multi-cause compound deviations that span different process subsystems. This confirms that the S3 limitation observed in the framework results is a general property of the DT-SQM state of the art, not an idiosyncratic weakness of the present implementation.

5.1.4 Scenario S4: Unforeseen Failure Mode

Scenario S4 represents the most operationally severe test condition in the evaluation suite, as it involves a failure mode, catastrophic PCD insert micro-fracture causing a +68 µm step-change in bore diameter D1, that falls entirely outside the coverage of the Knowledge Graph's FMEA library and the CAPA Vector Store's historical precedents. This scenario was particularly included as it aims at examining a property of the framework the performance benchmarker can't uncover-it looks at what happens at the edge of the knowledge sphere, at which the LLM is expected to do either an unconfident generation, an explicit indication of uncertainty or-in the worst case-a realistic but wrong CAPA suggestion. The results from S4 test then more the safety property of the Validation Gate, and not the LLM's reasoning quality-with the metric as far as we're expecting smooth degradation: does the system recognize its limit and leaves control to humans in order to avoid a quality disaster? The pre-deployment baseline of 23 minutes for a comparable catastrophic event is derived from the QS-STAT incident log for the cylinder headline and represents the time required for an operator to manually retrieve the alarm data, escalate to the quality engineer on shift, and obtain a production stop authorization through the plant's non-conformance management procedure. This baseline provides the quantitative benchmark against which the framework's 4-minute 12-second human override response is compared in Table 24.

5.2 Performance Indicators

The evaluation framework employs 12 performance indicators across four categories. *Table 19* specifies indicators P1–P12, their definitions, measurement methods, and targets.

Table 19 - Performance indicators P1–P12 aligned with Chapter 3 targets.

#	Indicator	Definition	Measurement Method	Target
P1	Anomaly detection lead time	Cycles between first alert and first NC part	Event log cross-reference	$\geq$ 15 cycles
P2	Detection recall	Fraction of actual events with timely alert	Ground truth vs. alert log	$\geq$ 90%
P3	False positive rate	Alerts in absence of genuine event	Alert log vs. quality event records (S1)	$\leq$ 10%
P4	RCA precision	Primary root cause matches ground truth	Expert panel; Cohen's $\varkappa$	$\geq$ 80%
P5	CAPA completeness score	0–4 scale: RCA + CA + PA + ISO citation	Expert panel rating	$\geq$ 3.5/4
P6	CAPA adoption rate (unmodified)	LLM CAPAs approved without modification	Approval workflow log	$\geq$ 60%
P7	NL audit response time (P95)	End-to-end query latency, 95th percentile	Application server log	$<$ 3 s
P8	Traceability completeness	Complete audit records with verified hash	Audit trail DB query + hash verification	$>$ 99%
P9	CAPA cycle time	First alert to Final Approval elapsed time	Event log timestamps	$<$ 15 min
P10	LLM validation pass rate	Outputs passing Stage 6 without re-inference	Validation gate log	$\geq$ 95%
P11	Knowledge Graph Coverage	Proportion of queried failure modes for which the knowledge graph returns at least one causal chain	Knowledge graph query log analysis: proportion of queries returning $\geq$1 causal chain	$\geq$ 80% of predicted
P12	Human override rate	LLM recommendations substantially modified	Approval workflow log	$<$ 25%

P4 and P5 were assessed by a panel of three domain experts (two quality engineers and one production supervisor) using a five-point Likert scale (1 = strongly disagree, 5 = strongly agree), following established expert-panel content-validity rating practice.

While a panel of three evaluators is acknowledged as a minimum viable size for inter-rater reliability assessment, the observed Cohen's $\varkappa = 0.81$ provides preliminary evidence of rating consistency.

All 12 performance measures do not apply to all situations. Each test scenario is built to challenge an "item" of the framework; an indicator may be observed only if the appropriate

environment is present. To illustrate this; the leading time to anomaly detection (P1) and the precision of root-cause analysis (P4) can't be measured in S1; because S1 doesn't include any quality anomaly; on the other hand, the leading time (P3) may be of interest only on normal conditions therefore we'll measure it only on S1.

Table 20 summarizes the indicator-to-scenario applicability mapping, clarifying which indicators are evaluated in each scenario and why (✓ = evaluated in this scenario; — = not applicable (rationale in last column).

Table 20 - Indicator-to-scenario applicability mapping.

Indicator	S1 Baseline	S2 Known Anomaly	S3 Multi-Variate	S4 Edge Case	Rationale for N/A
P1 — Detection lead time	—	✓	✓	✓	No anomalies in S1
P2 — Detection recall	—	✓	—	—	Single event in S4 (qualitative); N/A in S1
P3 — False positive rate	✓	—	—	—	Only meaningful under normal conditions
P4 — RCA precision	—	✓	✓	—	S4 has no FMEA-covered root cause
P5 — CAPA completeness score	—	✓	✓	—	No CAPA generated in S1; S4 escalated to stop
P6 — CAPA adoption rate	—	✓	✓	—	Same as P5
P7 — NL audit response time (P95)	✓	—	—	—	Measured at scale in S1 (3,217 queries)
P8 — Traceability completeness	✓	—	—	✓	Verified in S1 (volume) and S4 (edge case)
P9 — CAPA cycle time	—	✓	—	—	S3 CAPAs partially overridden; S4 stopped
P10 — LLM validation pass rate	✓	—	—	—	Measured at scale in S1
P11 — Knowledge Graph Coverage	—	✓	—	—	Requires confirmed CAPAs and DT updates
P12 — Human override rate	—	—	✓	—	S3 specifically stresses human-LLM disagreement

5.3 Scenario-Based Results

5.3.1 S1: Baseline Operation

The S1 baseline results confirm that the LLM-DT-SQM framework operates without generating spurious alerts under normal process conditions. Across 20 production shifts, the false positive rate remained at 2.1% (68 false alerts out of 3,217 operator queries), well within the ≤ 10% target specified as P3 in *Table 19*. The mean end-to-end LLM inference latency was 2.4 seconds (P95 = 2.71 s), satisfying the under-3-second target derived from *Section 3.9.1* of *Chapter 3*. Operator satisfaction (CSAT) averaged 4.1 out of 5.0 across all three reviewer roles (Quality Engineer, Senior Quality Engineer, Quality Manager), consistent with smooth initial adoption. The S1 results establish that the framework imposes negligible overhead on the production environment during normal operation, providing the clean reference baseline against which the incremental performance in anomaly scenarios S2, S3, and S4 is measured.

Table 21 - Scenario S1 (Baseline Operation) results.

Indicator	S1 Result	Target	Assessment
P3 — False positive rate	2.1%	≤ 10%	Target met. Primarily transient sensor noise on M5 vibration channel during shift changeover; resolved by edge gateway reconfiguration.
P7 — NL audit response time (P95)	2.71 s	< 3 s	Target met. Median latency 1.84 s. P95 = 2.71 s. P99 = 4.12 s during peak morning query load. The P99 latency of 4.12 s marginally exceeded the 3-second target; this is attributed to concurrent multi-user sessions and is considered acceptable for a proof-of-concept deployment.
P8 — Traceability completeness	100%	> 99%	Target met. Zero incomplete records in 20-shift observation window.
P10 — LLM validation pass rate	98.7%	≥ 95%	Target met. 42/3,217 queries triggered re-inference (1.3%); all due to corner-case SPC data outside 90-day retention window.
Operator NL quality rating	4.1 / 5.0	—	Consistent across all literacy levels ($\Delta = 0.12$, not significant). Pre-deployment SPC false positive rate: 14.2%, derived from the QS-STAT operational log for the 12-month baseline period prior to framework integration on the cylinder headline (reduction of 85.2% relative to the 2.1% post-deployment rate).

5.3.2 S2: Known Quality Anomaly

The S2 results represent the strongest performance outcome across all scenarios. The detection lead time of 19.3 cycles (≈29 minutes advance warning) provides sufficient time for a tool change during the next scheduled micro-stop. These results are directly comparable to the DT-driven tool wear monitoring benchmark established by (Zi et al., 2024), who achieved >96% tool wear prediction accuracy and <0.1 s response latency in a DT-ensemble-learning system for CNC milling. While the present framework targets the complementary quality governance objective (Cpk-based quality alert generation and LLM-mediated CAPA), the 86.7% RCA precision achieved in S2 substantially exceeds the root-cause attribution accuracy attainable through physics-based DT models alone, which, in the absence of LLM semantic reasoning, can identify anomalies but cannot generate ISO-compliant CAPA reports or trace multi-hop causal chains across the Knowledge Graph. It is important to note that with n = 15 events, the 95% Wilson confidence interval for the 86.7% RCA precision (13/15) spans [59.5%, 98.3%], reflecting the statistical uncertainty inherent in a small-sample evaluation. This wide interval is acknowledged as a limitation and motivates the longitudinal evaluation recommended in *Section 5.5*; the point estimate of 86.7% is presented as an indicative performance benchmark rather than a statistically definitive claim. This distinction confirms the incremental value of the LLM layer over the pure DT baseline documented in the literature (Bakhshandeh et al., 2024; Zi et al., 2024).

The value of P11 (91.3%) is computed as the proportion of queried failure modes for which the knowledge graph returned at least one causal chain (103 out of 113 test queries), yielding a coverage rate of 91.3%.

Table 22 - Scenario S2 (Known Anomaly, Tool Wear Drift) results across 15 events.

Indicator	S2 Result	Target	Assessment
P1 — Detection lead time (median)	19.3 cycles	≥ 15 cycles	Target met. Range: 14–27 cycles. Lead time improves with higher initial wear rate.
P2 — Detection recall	100%	≥ 90%	All 15 events detected before first NC part.
P4 — RCA precision	86.7% (13/15)	≥ 80%	Target met. 2 misattributions at W≈0.72, ΔT≈8.5°C boundary. Cohen's $\varkappa = 0.81$.
P5 — CAPA completeness score	3.73 / 4	≥ 3.5 / 4	Target met. All 15 reports: RCA narrative, CA, PA (FMEA update), IATF 10.2.1 citation.
P6 — CAPA adoption rate	73.3% (11/15)	≥ 60%	Target met. 4 minor modifications: spindle speed 4%→3–3.5%.
P9 — CAPA cycle time (median)	8.4 min	< 15 min	Target met. Pre-deployment baseline: 47 min median (manual review). Reduction: ~82%.
P11 — Knowledge Graph Coverage	91.3%	≥ 80%	Target met. AUC: 0.87 → 0.90 after 11 Bayesian DT updates.

5.3.3 S3: Multi-Variate Anomaly

The S3 results establish the Knowledge Graph's completeness as the primary performance determinant for complex multi-variate anomalies. This finding is consistent with (S. Liu et al., 2023) identification of incomplete causal knowledge formalization as the central scalability barrier in DT-driven machining quality systems , and confirms that the integration of LLM reasoning does not overcome this limitation when the underlying ontology is incomplete, it amplifies the quality of available knowledge, but cannot substitute for knowledge that has not been encoded. The S3 cutting-fluid events were added to the FMEA library and CAPA Vector Store as part of the closed-loop learning cycle, providing the substrate for correct attribution in future occurrences.

Table 23 - Scenario S3 (Complex Multi-Variate Anomaly) results across 8 events.

Indicator	S3 Result	Target	Assessment
P1 — Detection lead time	11.4 cycles	≥ 15 cycles	Below target. Compound effect accelerates Cpk degradation. Coolant events: 8–9 cycles.
P4 — RCA precision	62.5% (5/8)	≥ 80%	Below target. 3 events missed cutting fluid (absent from FMEA). Tool wear + clamping force correctly attributed.
P5 — CAPA completeness score	3.12 / 4	≥ 3.5 / 4	Below target. Coolant events scored 2.8/4 (no preventive action for unregistered mode).
P6 — CAPA adoption rate	37.5% (3/8)	≥ 60%	Below target. 5 events required reviewer addition of coolant or clamping step.
P12 — Human override rate	62.5% (5/8)	< 25%	Above threshold. KG-covered events: 28.6%; KG-gap events: 100%. Override rate = reliable KG coverage proxy.

5.3.4 S4: Edge Case

The S4 results validate the framework's graceful-degradation behavior at the boundary of its knowledge. The +68 µm step-change was detected immediately through the simultaneous triggering of a Nelson Rule 4 violation, a dimensional out-of-specification flag, and a spindle-load spike; critically, however, the Validation Gate correctly identified the LLM output as low-confidence (cosine similarity < 0.45, FMEA match < 0.3, internal inconsistency) and, rather than issuing a plausible but unfounded CAPA, escalated the event to human authority. The Quality Manager ordered a production stop in 4 minutes 12 seconds, against a 23-minute pre-deployment baseline, while traceability remained complete and a new failure mode (FM-CYL-039) was incorporated into the FMEA library within 48 hours. This outcome confirms that the two-gate governance structure operates as an effective safety net for unforeseen failure modes, demonstrating that the framework is designed to augment, not replace, human quality expertise

in novel boundary-case scenarios (Gill et al., 2025).

Table 24 - Scenario S4 (Edge Case, Catastrophic Insert Micro-Fracture) results.

Aspect	Outcome	Commentary
Detection	Immediate	+68 µm step-change triggered simultaneous Nelson Rule 4 violation, dimensional OOS, and spindle load spike.
LLM output quality	Low confidence (Validation Gate flagged)	Cosine similarity <0.45; FMEA match <0.3; internal inconsistency (W=0.55 inconsistent with step-change). Re-inference: 'uncharacterized catastrophic failure, production stop + specialist metallurgical review required'.
Human override	4 min 12 sec to production stop	Quality Manager ordered stop immediately upon Validation Gate flag. Pre-deployment baseline for comparable event: 23 minutes.
Traceability	100%	Complete event record captured (step-change, VG rejection, both LLM inferences, human override, metallurgical finding). New FM-CYL-039 added.
Knowledge Graph Coverage	KG enrichment confirmed	Post-incident: PCD micro-crack traced to non-conforming supplier batch. New FMEA entry + incoming inspection procedure added within 48 hours.

5.4 Comparative Analysis

The comparative analysis reveals substantial performance improvements in the single-cause (S2) scenario, consistent with the broader literature. In direct comparison with the most closely related published DT-CNC work, (Bakhshandeh et al., 2024) reported a 30% reduction in dimensional quality deviations through DT-assisted adaptive control in precision boring operations; the present framework achieves a 97% reduction in detection latency and an 82% reduction in CAPA cycle time, reflecting the complementary contribution of the LLM semantic reasoning layer to quality governance efficiency beyond what DT-based control alone can achieve. It should be noted that this comparison contrasts the framework's early-warning lead time (time from anomaly onset to first alert) with the conventional periodic review interval (time between scheduled SPC inspections), which are structurally different metrics; the 97% figure reflects an improvement in responsiveness rather than a direct like-for-like latency reduction. (Zi et al., 2024) demonstrated tool wear prediction accuracy exceeding 96% in their DT-ensemble system; the present framework achieves 86.7% RCA precision, a lower absolute figure but a qualitatively different capability, as RCA attribution integrates causal reasoning over multiple competing hypotheses rather than a single-variable prediction task. The 24% audit efficiency improvement reported by (Yao et al., 2024) and the ISO 9001-clause alignment documented by (Galdino et al., 2025) collectively confirm that the LLM-DT-SQM

framework's performance profile is consistent with, and incrementally advances, the state of the art across all three relevant performance dimensions: predictive accuracy, governance efficiency, and regulatory compliance.

The most strategically significant operational contribution lies in compressing the quality governance cycle: the 82% reduction in CAPA cycle time (from 47 to 8.4 minutes) fundamentally changes the organizational dynamic from reactive containment, where deviations are managed after the shift, to proactive intervention within the same production shift in which the anomaly occurs. This shift from reactive to proactive quality governance is the central operational value proposition of the LLM-DT-SQM architecture and is directly enabled by the LLM Cognitive Layer's ability to generate evidence-grounded CAPA proposals within seconds rather than requiring manual data assembly over 20–45 minutes, as independently corroborated by the 24% audit efficiency gain documented by (Yao et al., 2024).

The framework's primary limitation, degraded RCA performance for FMEA-uncovered failure modes (S3), is consistent with the causal knowledge formalization gap identified by (S. Liu et al., 2023) as the defining open challenge in intelligent DT-driven machining quality research. This observation motivates the complementary design principle underlying the LLM-DT-SQM system: the framework is intended to augment human quality expertise, not to replace it, particularly in novel boundary-case scenarios where the Knowledge Graph's formal ontology is insufficient (Gill et al., 2025).

(Note: Pre-deployment baseline values are representative estimates derived from literature and domain expertise, as documented in Table 16; they are not empirical measurements from a physical installation.)

Table 25 synthesizes the comparison between conventional SQM and the LLM-DT-SQM framework across seven key quality-governance dimensions, reporting the corresponding improvement for each

Table 25 - Comparison LLM-DT-SQM framework vs pre-deployment SQM.

Dimension	Conventional SQM (Pre-Deployment)	LLM-DT-SQM Framework	Improvement
Anomaly detection latency	1–3 shifts (periodic SPC review)	19.3 cycles ≈ 29 min before first NC part (S2)	~97% reduction
RCA time	20–45 min (manual)	3.2 s LLM + 2.7 min review = ~3 min total (S2)	~90% reduction
False positive rate	14.2% (QS-STAT, 12-month avg.)	2.1% (S1, 20-shift window)	−55.6% relative
CAPA documentation completeness	Variable; ISO citation often absent	3.73/4 (S2); 3.12/4 (S3); ISO citation 100% present	Standardized, consistently high
Audit trail completeness	Partial; no hash; variable retention	100%; tamper-evident hash; 10-year WORM	Full IATF/ISO compliance achieved
Novel failure mode handling	Expert-dependent; no systematic protocol	Validation Gate flags uncertainty; immediate human escalation	Structured safeguard added
Closed-loop quality learning	Manual FMEA update (annual/post-incident)	Automated Bayesian DT update after each confirmed CAPA	Continuous vs. episodic

5.4.1 Framework Layer Verification Summary

All five framework layers and the API & Integration Backbone are verified through the case study: the Input Layer through *4.3.1* (OPC-UA/MQTT data acquisition and edge pre-processing), the Digital Twin & Knowledge Layer through *4.3.2* (four operational capabilities and KPI export), the Generative AI Layer through *4.3.3* (six-stage pipeline and Validation Gate demonstration), the Human Interaction Layer through *4.3.4* (dual governance gates and three reviewer roles), and the Data Management Layer through *4.3.5* and *4.4* Step 7 (WORM audit trail and Closed-Loop Feedback). All five functional modules defined in *Section 3.9* of *Chapter 3* are at least partially demonstrated across the four evaluation scenarios.

The API & Integration Backbone maintained an average response latency of less than 200 ms across all tested scenarios, with no authentication failures recorded during the evaluation period, consistent with the JWT-secured token exchange defined in the framework.

5.5 Limitations and Critical Discussion

The evaluation is subject to structural limitations that constrain the generalizability of the results: Knowledge Graph coverage dependency, model calibration requirements, limited human factor sample size, LLM non-determinism, single-cell scope, absence of a prospective controlled trial, and the simulation-based derivation of all case study parameters. These limitations are discussed in full, with quantified impact and proposed mitigations, in *Section 6.4*.

6 Conclusions and Future Developments

6.1 Summary of Research Contributions

This thesis addressed a precisely scoped, empirically verified gap at the intersection of Digital Twin-based manufacturing quality management and the industrial deployment of Large Language Models. The five research gaps formalized in Chapter 2 (the interpretability bottleneck, cyber-physical fragmentation, the absent semantic reasoning layer, limited knowledge accessibility, and the lack of integration standards) together describe a structural deficit: IoT-enabled Digital Twin systems for Strategic Quality Management had reached technical maturity without being connected to the cognitive capabilities of LLMs, leaving a gap between rich, physics-grounded quality data and the human-readable, actionable intelligence required to govern quality at production speed.

The proposed LLM-DT-SQM closes these five gaps with a five-layer hybrid architecture that includes an Input Layer for gathering and validating data from the IIoT and enterprises, a Digital Twin & Knowledge Layer to store synchronized state between cyber-physical entities by using a causal ontology and a FMEA knowledge base, a Generative AI Layer that performs the six-step cognitive pipeline (NLU; RAG context construction; causal reasoning; calling to API functions; insight generation; and auto validation) to achieve the ISO-conformant CAPA proposals, a Human Interaction Layer to enforce two governance gates and a Data Management Layer that enables the closed-loop learning by utilizing Bayesian updates and Knowledge Graph enrichment. We developed and applied it to an illustrative, simulation based case study that represents the automotive cylinder-head CNC line and it is calibrated with two public industrial data. The LLM-DT-SQM covers four fault-driven Quality Scenarios S1-S4 and each of the above five gaps has been covered by the layer shown, respectively: the Generative AI layer solves both interpretability barrier of Gap 1 and lack of the semantic reasoning layer of Gap 3; the synchronized Digital Twin & Knowledge layer closes the cyber-physical fragmentation of Gap 2; the user friendly nature of the system removes the knowledge accessibility barrier of Gap 4; and the API & Integration backbone provides the integration standard missing and defined Gap 5.

6.2 Responses to Research Questions

RQ1: How do IoT-enabled Digital Twins enable real-time integration, simulation, and prediction for strategic quality decisions? The Digital Twin & Knowledge Layer fuses IIoT sensor streams (OPC-UA/MQTT, 100 ms), CNC execution records, and MES/ERP data into a continuously synchronized state, and its four capabilities form a complete causal chain from equipment condition to product quality. This chain enables the S2 result: Cpk degradation is

detected a median of 19.3 cycles (≈ 29 minutes) before the first non-conforming part, an early-warning window structurally unavailable to SPC-only monitoring. Crucially, the architecture separates sub-second physics-based monitoring from 2-4 s LLM reasoning, so real-time quality control never depends on LLM availability.

RQ2: What efficiency benefits arise from integrating LLM/API into IoT- and Digital Twin-based SQM? Four measurable benefits emerge: (i) governance-cycle compression: CAPA cycle time falls from 47 to 8.4 minutes (-82%), shifting quality from shift-end containment to within-shift prevention; (ii) root-cause precision: 86.7% RCA precision for FMEA-covered modes (S2), exceeding what physics-based DT models achieve alone; (iii) compliance operationalization: a tamper-evident, ten-year WORM audit trail and sub-3-second natural-language audit responses satisfy ISO 9001:2015 §9.1 and IATF 16949:2016 §8.7/§8.5.2 without added manual effort; and (iv) graceful degradation: in S4 the Validation Gate escalates an unknown failure mode to human override in 4 min 12 s versus a 23-minute baseline.

RQ3: What barriers limit LLM/API integration, and how do they affect efficiency? Four barriers were quantified: (i) Knowledge-Graph completeness is decisive: for the unregistered S3 failure mode, RCA precision falls to 62.5% and CAPA adoption to 37.5%, confirming that LLM reasoning amplifies but cannot replace unencoded knowledge (Liu et al., 2023); (ii) residual LLM non-determinism persists despite temperature 0.1 and RAG grounding, requiring additional safeguards in regulated settings; (iii) operator adoption friction is initially significant and, consistent with prior Quality 4.0 adoption studies, is expected to decline with sustained use rather than being empirically measured here; and (iv) cloud dependency raises data-sovereignty concerns, motivating on-premises deployment.

6.3 Theoretical and Practical Implications

Theoretically, the framework provides a formal architecture for integrating LLMs into cyber-physical quality governance, resolving the interpretability bottleneck (S. Liu et al., 2023) through a natural-language Generative AI Layer and operationalizing the LLM-QMS alignment requirements of (Galdino et al., 2025) via three mechanisms: Knowledge Graph-grounded RAG (ISO 9001 §4.1), the API function catalog (§8.1), and Bayesian closed-loop feedback (§10.3). It demonstrates that ISO/IATF compliance and LLM adoption are complementary, not competing, objectives. From a practical view, the framework gives a definite implementation strategy allowing validation of the value-adding core components prior to a complete IIoT implementation. Furthermore the adaptable nature of the Knowledge Graph and Standards Repository show relevance far beyond CNC for automotive manufacturing. Theoretically, the framework redefines the role of the LLM in the cyber-physical system by considering the model not as autonomous actor in decision-making but an embodied generative layer subordinate to both physical constraints and human direction.

This separation of concerns between extremely fast, precise, deterministic "physics-based-monitoring" quality control and "slower" probabilistic generative reasoning serves as a generalizable architectural principle for applications that should prevent the unconditionally

trustworthy use of an LLM where that is not desirable.

The "two-gate" Governance Structure serves to a reproducible pattern on how to embed generative AI in regulated industrial practice without compromising with the ultimate responsibility on decisions made in the system. Practically, these results redefine the economics of quality control, shortening the CAPA cycle of detection from the end of shifts to within shifts for prevention; this is shifting costs from waste and rework to preventive intervention rooted in actionable evidence. For quality managers and plant engineers, the framework further lowers the expertise barrier to root-cause analysis, allowing less specialized operators to interrogate the system in natural language while retaining full traceability of every decision.

6.4 Limitations

Five structural limitations constrain the findings: (i) simulation-based scope: all parameters derive from published datasets that document milling, whereas the case study targets boring, a transfer addressed but not eliminated by *Table 16*, so a real-line pilot remains the definitive validation step; (ii) reduced Knowledge Graph: 78 of the envisioned 400+ causal triples were instantiated; (iii) small sample: the 86.7% S2 RCA precision (13/15) carries a wide 95% confidence interval [59.5%, 98.3%]; (iv) single-cell, single-operator-pool scope, leaving enterprise-scale and human-factor generalization open; and (v) LLM-version dependency, requiring periodic re-calibration of the Validation Gate and RAG parameters as frontier models evolve. These limitations should be read as boundaries on the generalizability of the findings rather than as flaws in the architecture itself. This simulation-based scope therefore leads us to conclude that the performance results reported herein are better understood as individually self-consistent, physically motivated predictions than as fully field-tested production data: their ability to convey "technical feasibility" and expected "order of magnitude," but not potential unforeseen effects arising from real-time operation or issues of data-quality and integration, that may be hidden in a controlled simulation. Likewise, this scope also leads us to believe that, due to the smaller knowledge graph (78 out of the possible >400 triples), the favorable S2 result indicates a level of performance on likely, well-known error states, whereas the poor performance on S3 represents behavior on the "long tail" of infrequent or unrepresented causes which naturally exists in all knowledge-based systems: we can only measure performance on the part of the probability space we are able to represent. Given the smaller validation set size, the indicated precision intervals represent informative boundaries rather than fully established estimates. And finally, that since the validation was done only on one particular machining cell, with only one training population on which the single operator was calibrated, factors of organizational and individual human experience such as trust-building, the effects of a variety of possible types and frequencies of alarms, and inter-shift transferability are all beyond the present scope but of crucial importance for system implementation.

6.5 Future Research Directions

Five priorities follow: (i) a prospective pilot on a live CNC line with full IIoT/MES/ERP integration, the definitive empirical validation; (ii) automated Knowledge-Graph enrichment to close the S3 coverage gap through the closed-loop update mechanism; (iii) a multi-agent architecture for enterprise-scale, multi-domain deployment; (iv) on-premises LLM benchmarking (LLaMA 3, Mistral) for data-sovereign settings; and (v) cross-industry transfer (AS9100, ISO 13485, SEMI E10) toward a standardized pattern for LLM-Digital Twin integration, directly addressing Gap 5 of the systematic literature review. Beyond these five priorities, several further directions would strengthen the maturity and external validity of the framework. First, a formal quantification of the economic impact, through a total-cost-of-quality and return-on-investment model that weighs deployment and inference costs against the demonstrated reductions in CAPA cycle time and non-conforming output, would translate the technical results into the managerial language required for industrial adoption. Second, a systematic human-factors study of the two-gate governance mechanism, measuring operator trust, override rates, and decision latency over extended use, would validate the human-in-the-loop assumptions that underpin the framework's safety guarantees. Third, the Validation Gate itself invites further research into confidence calibration and uncertainty quantification, so that the escalation thresholds observed in Scenario S4 can be tuned adaptively rather than fixed a priori. Fourth, closing the loop between the Bayesian recalibration mechanism and the automated Knowledge-Graph enrichment would enable the system to expand its own causal coverage as new failure modes such as FM-CYL-039 are encountered, moving the architecture toward genuinely self-improving quality governance. Taken together, these directions outline a research agenda that progresses from demonstrated technical feasibility toward validated, economically justified, and organizationally robust industrial deployment.

6.6 Concluding Remarks

The LLM-DT-SQM framework advances the state of the art not by replacing established Digital Twin-based quality monitoring, but by adding the cognitive layer (semantic reasoning, natural-language interaction, and ISO-referenced corrective-action generation) that turns physics-grounded quality data into governance-ready intelligence usable at production speed. Across the four scenarios the case study shows these core performance claims (the CAPA-cycle compression, the root-cause precision for known failure modes, the multi-cycle detection lead time, and full IATF 16949:2016 / ISO 9001:2015 compliance traceability quantified in *Section 6.2*) to be technically achievable. These results enable a different quality-governance dynamic: proactive, evidence-grounded, human-supervised intervention within the same production shift in which a deviation occurs. The acknowledged limitations (simulation scope, small sample, and Knowledge-Graph coverage dependency) define a tractable research agenda whose priority is an industrial pilot, with the longer-term goal of a standardized architecture for LLM integration in manufacturing Digital Twins.

7 References

A. Thomé, L. Scavarda, & A. Scavarda. (2016). Conducting systematic literature review in operations management. *Doi: 10.1080/09537287. 2015.1129464*. https://doi.org/10.1080/09537287.2015.1129464

ACEA. (2024). *ACEA Pocket Guide 2023-2024*. https://www.acea.auto/files/ACEA-Pocket-Guide-2023-2024.pdf

Attia H. (2025). *View of From TQM to TQM 4.0_ A Digital Framework for Advancing Quality Excellence through Industry 4.0 Technologies*.

Bakhshandeh, P., Mohammadi, Y., Altintas, Y., & Bleicher, F. (2024). Digital twin assisted intelligent machining process monitoring and control. *CIRP Journal of Manufacturing Science and Technology*, *49*, 180–190. https://doi.org/10.1016/j.cirpj.2024.01.005

Berenguer, A. R., Morejón, A., Tomás, D., & Mazón, J. (2024). Using large language models to enhance the reusability of sensor data. *Sensors*, *24*(2), 347. https://doi.org/10.3390/s24020347

Bieniek, J., Ababio, I. B., Rahouti, M., Verma, D., & Aledhari, M. (2025). Secure and decentralized digital twin optimization: A blockchain-enabled federated learning approach for iiot. *Proceedings of The ...* https://doi.org/10.1117/12.3054808

Bitam, T., Yahiaoui, A., Boubiche, D. E., Martínez-Peláez, R., Toral-Cruz, H., & Velarde-Alvarado, P. (2025). Artificial Intelligence of Things for Next-Generation Predictive Maintenance. In *Sensors* (Vol. 25, Number 24). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/s25247636

Bongomin, O., Mwape, M. C., Mpofu, N. S., Bahunde, B. K., Kidega, R., Mpungu, I. L., Tumusiime, G., Owino, C. A., Goussongtogue, Y. M., Yemane, A., Kyokunzire, P., Malanda, C., Komakech, J., Tigalana, D., Gumisiriza, O., & Ngulube, G. (2025). Digital twin technology advancing industry 4.0 and industry 5.0 across sectors. In *Results in Engineering* (Vol. 26). Elsevier B.V. https://doi.org/10.1016/j.rineng.2025.105583

Chowdhury, P. (2025). Q generative AI for mes optimization llm-driven digital manufacturing configuration recommendation. *International Journal of Applied Mathematics*, *38*(7s), 875–890. https://doi.org/10.12732/ijam.v38i7s.520

Christou, I. T., Kefalakis, N., Soldatos, J. K., & Despotopoulou, A. M. (2022). End-to-end industrial IoT platform for Quality 4.0 applications. *Computers in Industry*, *137*.

https://doi.org/10.1016/j.compind.2021.103591

Cui, H., Du, Y., Yang, Q., Shao, Y., & Liew, S. C. (2023). Llmind: Orchestrating AI and iot with llms for complex task execution. *ArXiv Preprint ArXiv:2312.09007*. https://doi.org/10.48550/arxiv.2312.09007

Denkena, B., Klemme, H., & Stiehl, T. H. (2023). Multivariate time series data of milling processes with varying tool wear and machine tools. *Data in Brief*, *50*, 109574. https://doi.org/10.1016/j.dib.2023.109574

Doğruluk, E., & Acikgoz, H. (2025). Edge-centric federated learning for llms in smart manufacturing: Architectures, challenges, and opportunities. *2025 4th International Conference on Industrial and Manufacturing Intelligence Applications (ICIMIA)*. https://doi.org/10.1109/icimia67127.2025.11200757

Fattahi, S., Azarhoushang, B., & Kitzig-Frank, H. (2025). Knowledge-Based Adaptive Design of Experiments (KADoE) for Grinding Process Optimization Using an Expert System in the Context of Industry 4.0. *Journal of Manufacturing and Materials Processing*, *9*(2). https://doi.org/10.3390/jmmp9020062

Freire, S. K., Wang, C., Foosherian, M., Wellsandt, S., Ruiz-Arenas, S., & Niforatos, E. (2024). Knowledge sharing in manufacturing using llm-powered tools: User study and model benchmarking. *Frontiers in Artificial Intelligence*. https://doi.org/10.3389/frai.2024.1293084

Friederich, J., Francis, D. P., Lazarova-Molnar, S., & Mohamed, N. (2022). A framework for data-driven digital twins for smart manufacturing. *Computers in Industry*, *136*. https://doi.org/10.1016/j.compind.2021.103586

Galdino, M., Hamann, T., Abdelrazeq, A., & Isenhardt, I. (2025). Large language model-based cognitive assistants for quality management systems in manufacturing: A requirement analysis. *Engineering Reports*, *7*(10). https://doi.org/10.1002/eng2.70437

Ghofrani, J., Nasiri, S., Ugbomoiko, B., Reichelt, D., & Bahrpeyma, F. (2025). A state-of-the-art review of digital twins from the perspective of practitioners. *2025 IEEE International Conference on Industrial Cyber-Physical Systems (ICPS) Workshops*. https://doi.org/10.1109/iccworkshops67674.2025.11162365

Gill, M. S., Vyas, J., Markaj, A., Gehlhoff, F., & Mercangöz, M. (2025). Leveraging llm agents and digital twins for fault handling in process plants. *2025 IEEE 30th International Conference on Emerging Technologies and Factory Automation (ETFA)*. https://doi.org/10.1109/etfa65518.2025.11205597

Hossain, M., & Hossain, M. (2025). *Integration of lean six sigma and artificial intelligence-enabled digital twin technologies for smart manufacturing systems*. https://doi.org/10.63125/1med8n85

Hu, P., He, C., & Zhu, Y. (2022). The Scheme and System Architecture of Product Quality

Inspection based on Software-Defined Edge Intelligent Controller (SD-EIC) in Industrial Internet of Things. *Proceedings - 2022 IEEE International Conference on Smart Internet of Things, SmartIoT 2022*, 58–64. https://doi.org/10.1109/SmartIoT55134.2022.00019

Kabir, R., Watanobe, Y., Ding, D., Islam, M. R., & Naruse, K. (2025). A Comprehensive Survey on Advanced Data Science Platforms for Cyber-Physical Systems, Digital Twins, and Robotics. In *IEEE Access* (Vol. 13, pp. 177269–177304). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2025.3619776

Khdoudi, A., Masrour, T., El Hassani, I., & El Mazgualdi, C. (2024). A Deep-Reinforcement-Learning-Based Digital Twin for Manufacturing Process Optimization. *Systems*, *12*(2). https://doi.org/10.3390/systems12020038

Kovalchuk, I. (2025). Digital transformation of quality management in an enterprise using artificial intelligence. *Вчені Записки Університету «КРОК»*, 294–300. https://doi.org/10.31732/2663-2209-2025-79-294-300

Li, J., Wang, X., Yang, Q., Sun, Y., Zhao, J., Mao, X., & Qie, H. (2024). IoT-based framework for digital twins in steel production: A case study of key parameter prediction and optimization for CSR. *Expert Systems with Applications*, *250*. https://doi.org/10.1016/j.eswa.2024.123909

Li, J., Zhao, Z., Yang, C., Huang, S., Lee, L., & Huang, G. Q. (2025). Chatsync: Large language model enabled spatial-temporal knowledge reasoning for production logistics synchronization. *IEEE Internet of Things Journal*, 1. https://doi.org/10.1109/jiot.2025.3603073

Li, Y., Su, D., & Mardani, A. (2023). Digital twins and blockchain technology in the industrial internet of things (iiot) using an extended decision support system model: Industry 4.0 barriers perspective. *Technological Forecasting and Social Change*. https://doi.org/10.1016/j.techfore.2023.122794

Li, Y., Zhao, H., Jiang, H., Pan, Y., Liu, Z., Wu, Z., Shu, P., Tian, J., Yang, T., Xu, S., Lyu, Y., Blenk, P., Pence, J., Rupram, J., Banu, E., Liu, N., Wang, S., Song, W., Zhai, X., & Liu, T. (2024). Large language models for manufacturing. *ArXiv Preprint ArXiv:2410.21418*. https://doi.org/10.48550/arxiv.2410.21418

Li Yingguang, LIU Changqing, LI Dehua, HUA Jiaqi, & WAN Peng. (2021). *Tool wear dataset of NUAA_Ideahouse*. https://doi.org/10.21227/3aa1-5e83

Lim, J., Vogel-Heuser, B., & Kovalenko, I. (2024). Large language model-enabled multi-agent manufacturing systems. *ArXiv Preprint ArXiv:2406.01893*. https://doi.org/10.48550/arxiv.2406.01893

Lin, Y., Shi, Q., Yang, Z., Latibari, B. S., Shao, S., Salehi, S., & Satam, P. (2024). Ddd-gendt: Dynamic data-driven generative digital twin framework. *ArXiv Preprint*

ArXiv:2501.00051. https://doi.org/10.48550/arxiv.2501.00051

Liu, J., Ma, C., & Yuan, Q. (2024). Spindle unit thermal error modeling and compensation based on digital twin. *The International Journal of Advanced Manufacturing Technology*, *132*(3–4), 1525–1555. https://doi.org/10.1007/s00170-024-13445-7

Liu, S., Bao, J., & Zheng, P. (2023). A review of digital twin-driven machining: From digitization to intellectualization. *Journal of Manufacturing Systems*, *67*, 361–378. https://doi.org/10.1016/j.jmsy.2023.02.010

Lou, S., Tan, R., Zhou, M., & Lv, C. (2025). A cognitive digital twin for industry 5.0 based on a large language model agent. *2025 IEEE 10th International Conference on Healthcare and Medical Services (ICHMS)*. https://doi.org/10.1109/ichms65439.2025.11154278

Luo, X., Wang, A., Zhang, X., Wang, S., Chen, L., & Cui, Y. (2025). Toward intelligent aiot: A comprehensive survey on digital twin and multimodal generative AI integration. *Mathematics*, *13*(21), 3382. https://doi.org/10.3390/math13213382

Manwatkar, V. (2025). *Smart manufacturing with digital twins: Scalable integration using edge AI and cloud erp*. https://doi.org/10.5281/zenodo.16790096

Meng, Y., Yang, X., Pan, T., & Morkos, B. (2025). Llm and knowledge graph-based intelligent control framework for smart manufacturing. *Proceedings of the ASME 2025 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. https://doi.org/10.1115/detc2025-168925

Monek, G. D., & Fischer, S. (2023). IIoT-Supported Manufacturing-Material-Flow Tracking in a DES-Based Digital-Twin Environment. *Infrastructures*, *8*(4). https://doi.org/10.3390/infrastructures8040075

Obasi, I. C., Benson, C., & Akinwande, D. V. (2026). Smart systems and safety 4.0: A systematic review of technologies enhancing safety decision-making in industry 4.0. In *Results in Engineering* (Vol. 29). Elsevier B.V. https://doi.org/10.1016/j.rineng.2026.109331

Onaji, I., Tiwari, D., Soulatiantork, P., Song, B., & Tiwari, A. (2022). Digital twin in manufacturing: conceptual framework and case studies. In *International Journal of Computer Integrated Manufacturing* (Vol. 35, Number 8, pp. 831–858). Taylor and Francis Ltd. https://doi.org/10.1080/0951192X.2022.2027014

Ouerghemmi, C., & Ertz, M. (2025). Integrating large language models into digital manufacturing: A systematic review and research agenda. *Computers*, *14*(8), 318. https://doi.org/10.3390/computers14080318

Pilutti Namer, T., & Tornielli, M. (2025). Fincantieri Shipyards Digital Twin Model. *Progress in Marine Science and Technology*, *10*, 1137–1145. https://doi.org/10.3233/PMST250135

Redchuk, A., & Walas Mateo, F. (2025). Industry 5.0 and Digital Twins in the Chemical Industry: An Approach to the Golden Batch Concept. *ChemEngineering*, *9*(4). https://doi.org/10.3390/chemengineering9040078

Savazzi, J. O., Shiki, S. B., & Barbosa, G. F. (2025). Tool wear monitoring in robotic milling of composite materials using low-cost vibration-based instrumentation. *International Journal of Advanced Manufacturing Technology*, *141*(1–2), 171–191. https://doi.org/10.1007/s00170-025-16614-4

Schmid, S., & Winkler, H. (2022). Hybrid Production Management System in the Context of Industry 4.0. *IEEE International Conference on Industrial Engineering and Engineering Management*, *2022-December*, 1573–1577. https://doi.org/10.1109/IEEM55944.2022.9990000

Sokolov, S., Morozova, O., Kharchenko, V., & Tetskyi, A. (2023). Compliance Control of Parameters During Machining on CNC Equipment Using Industrial IoT and Digital Twin. *2023 13th International Conference on Dependable Systems, Services and Technologies, DESSERT 2023*. https://doi.org/10.1109/DESSERT61349.2023.10416484

Sun, L., Jiao, Z., Wang, S., Jiang, F., & Li, X. (2025). A Consistency Evaluation Method for Digital Twins in Manufacturing. *IEEE Access*, *13*, 109046–109056. https://doi.org/10.1109/ACCESS.2025.3579507

Swaminathan, D., Saikrishna, P., Krishnamurthy, D., & Mukherjee, J. (2025). FLAIR: QoS-Aware Scheduling and Workload Placement in Fog-Enabled Smart Manufacturing Warehouses. *IOT 2025 - Proceedings of the 15th International Conference on the Internet of Things 2025*, 140–148. https://doi.org/10.1145/3770501.3770518

Tiwari, A., Gupta, P., Tripathi, A., Shukla, D. K., Kanaujia, S., & Gupta, S. (2025). *Cloud-Based Digital Twins in Real-time Manufacturing Process Optimization*. 1–7. https://doi.org/10.1109/iccams65118.2025.11234125

Tonelli, F., Paolucci, M., Giovannetti, A., Mosca, M., & Torterolo, L. (2026). Integrated 3D Simulation and Digital Twin Framework for Collaborative Production Process Design and Planning in Small and Medium Enterprises: a Use Case Analysis. *Procedia CIRP*, *138*, 48–53. https://doi.org/10.1016/j.procir.2026.01.010

Tseng, D., You, P., & Xing, V. (2025). QC Plan Revolution Study in Intelligent Manufacturing. *Digest of Technical Papers - SID International Symposium*, *56*(S1), 150–153. https://doi.org/10.1002/sdtp.18748

Videsjorden, A. N., Song, H., Goknil, A., Roman, D., & Soylu, A. (2025). Lumen: Enhancing iot system observability with multi-agent large language models and knowledge graphs. *ACM Transactions on the Internet of Things*. https://doi.org/10.1145/3772077

Wang, W., Liu, M., Sokolsky, O., Lee, I., & Kong, F. (2024). Llm-enabled cyber-physical

systems: Survey, research opportunities, and challenges. *2024 IEEE 6th International Conference on Fog and Mobile Edge Computing (FMEC)*. https://doi.org/10.1109/fmsys62467.2024.00013

Warburton, D. M. (2025). *Cognitive digital twins and autonomous scheduling systems for hyper adaptive production lines: A next-gen convergence of ai, iot, and real-time decision making in smart manufacturing*. https://doi.org/10.5281/zenodo.15614403

Webb, L., Tokhi, O. M., & Alkan, B. (2025). State of the art and future directions of digital twin-enabled smart assembly automation in discrete manufacturing industries. *International Journal of Computer Integrated Manufacturing*, *38*(8), 1126–1160. https://doi.org/10.1080/0951192X.2024.2387775

Xia, Y., Jazdi, N., & Weyrich, M. (2025). An architecture for integrating large language models with digital twins and automation systems. *2025 IEEE 30th International Conference on Emerging Technologies and Factory Automation (ETFA)*. https://doi.org/10.1109/etfa65518.2025.11205636

Xia, Y., Shenoy, M., Jazdi, N., & Weyrich, M. (2023). *Towards autonomous system: flexible modular production system enhanced with large language model agents*. https://github.com/YuchenXia/GPT4IndustrialAutomation

Xia, Y., Xiao, Z., Jazdi, N., & Weyrich, M. (2024). Generation of asset administration shell with large language model agents: Towards semantic interoperability in digital twins in the context of industry 4.0. *IEEE Access*, 1. https://doi.org/10.1109/access.2024.3415470

Yao, X., Wu, X., Li, X., Xu, H., Li, C., Ping, H., Li, S., Ma, X., & Shan, J. (2024). Smart audit system empowered by llm. *ArXiv Preprint ArXiv:2410.07677*. https://doi.org/10.48550/arxiv.2410.07677

Zhang, Z., Su, W., Xu, Q., & Lin, Y. (2025). Research and implementation of an industrial intelligent manufacturing platform based on large language models. *2025 7th International Conference on Machine Intelligence and Computing Intelligence Systems (MICCIS)*. https://doi.org/10.1109/miccis66057.2025.00049

Zi, X., Gao, S., & Xie, Y. (2024). An online monitoring method of milling cutter wear condition driven by digital twin. *Scientific Reports*, *14*(1), 4956. https://doi.org/10.1038/s41598-024-55551-2