

An auditable large language model- and retrieval-augmented generation digital quality twin supporting simulation-based validation and safer defect control in Industry 4.0 and 5.0 for production and technology management

Jakub Pizon^{1*}, Łukasz Kański², Jan Chadam², Arkadiusz Gola³

¹ Department of Organisation of Enterprise, Faculty of Management, Lublin University of Technology, ul. Nadbystrzycka 38, 20-618 Lublin, Poland

² Department of Information Systems and Logistics, Faculty of Economics, Institute of Management and Quality Sciences, Maria Curie-Skłodowska University, Plac Marii Curie-Skłodowskiej 5, 20-031 Lublin, Poland

³ Department of Informatization and Robotization of Production, Lublin University of Technology, ul. Nadbystrzycka 36, Lublin, Poland

* Corresponding author's e-mail: j.pizon@pollub.pl

ABSTRACT

Generative artificial intelligence (GenAI), particularly large language models (LLMs), is increasingly considered for supporting production and technology management. In manufacturing quality engineering, LLMs may reduce the cognitive cost of defect analysis, accelerate root-cause diagnostics, and support quality documentation. Still, they can also generate plausible recommendations that are not grounded in evidence. This creates risks of erroneous process interventions, loss of traceability, and non-compliance with artificial intelligence governance requirements. This paper proposes an auditable LLM-augmented digital quality twin (LDQT) architecture that combines an LLM, Retrieval-Augmented Generation (RAG), a defect-centric digital quality twin, and human-in-the-loop approval. The architecture is evaluated in a discrete-event simulation of a production stream with stochastic process-drift events and three intervention policies: manual, rule-based, and LDQT. The revised computational study includes Monte Carlo analysis, explicit hypothesis testing, and sensitivity analysis. Under the adopted assumptions of 400 replications and 4000 items per replication, the LDQT scenario indicates an approximately 20% lower mean unit cost of quality and downtime than the manual baseline, as well as a higher first-pass yield. These findings should be interpreted as simulation-based evidence of architectural potential, not as proof of practical effectiveness in a real factory. The contribution is methodological: the study defines the Digital Quality Twin concept and provides metrics for balancing efficiency, auditability, and decision safety in industrial GenAI deployments.

Keywords: generative artificial intelligence, large language models, retrieval-augmented generation, digital quality twin, Industry 4.0, Industry 5.0, human-AI teaming, discrete-event simulation.

INTRODUCTION

The digital transformation of industry, commonly associated with Industry 4.0, remains uneven. Organisations most often invest in robotics/automation and machine connectivity, and only subsequently in data analytics and artificial

intelligence. The distribution of perceived relevance of individual technologies is illustrated in Figure 1, prepared based on data from a survey on the development of Industry 4.0 components in industrial enterprises [1]. For quality engineering, this has a fundamental consequence: process data are increasingly collected by sensors, supervisory

control and data acquisition (SCADA) systems, and manufacturing execution systems (MES), whereas the ability to interpret them and translate them into quality decisions frequently becomes the organisational bottleneck.

Industry 4.0 implementations are further constrained by resource and competence barriers. As shown in Figure 2, the most frequently cited challenges are time and funding shortages, topic complexity, and a lack of digital skills. From the perspective of designing GenAI solutions, this implies the need to build tools that not only improve key performance indicators (KPIs) but also reduce the entry barrier (time needed for analysis, access to expert knowledge) and minimise integration costs within existing information technology/operational technology (IT/OT) landscapes.

Against this backdrop, interest in generative AI has been growing rapidly. Data on planned deployments indicate that organisations initially choose language generative models, and only later embedding, multimodal, or diffusion models (Figure 3) [3]. The dominance of LLMs is understandable, as they provide a natural interface to knowledge dispersed across quality documentation, including failure mode and effects analysis (FMEA), control plans, eight disciplines (8D) reports, standard operating procedures (SOPs), and event logs. At the same time, the conversational layer alone does not solve the credibility problem. What is crucial in industrial applications is grounding responses in evidence and the ability to audit the decision process.

Contemporary research addresses this need by developing Retrieval-Augmented Generation (RAG), i.e., architectures that combine generation

with retrieval from external knowledge bases. The canonical RAG concept was formulated for knowledge-intensive tasks [4] and later expanded toward industrial systems, including support for manufacturing quality control [5]. RAG does not eliminate the risk of error, but it changes its nature. Instead of relying solely on a model's parametric knowledge, the system can indicate the sources underpinning a recommendation, thereby facilitating validation by a quality engineer.

In parallel, the Industry 5.0 paradigm is gaining relevance, emphasising human-centricity, sustainability, and resilience of production systems. In the European view, this entails shifting the emphasis from maximising automation to designing human-technology collaboration and process resilience to disruptions [6]. From this standpoint, LLMs should not be treated as a substitute for experts but as a tool that amplifies their work and shortens the time from signal to decision.

A separate yet critical axis is governance and compliance. In 2024, the European Union adopted the AI Act, introducing a risk-based approach and requirements for transparency and risk management for selected classes of systems [7]. In parallel, the International Organisation for Standardisation/International Electrotechnical Commission standard ISO/IEC 42001 supports the standardisation of AI management [8]. At the same time, the practical risk-management instrument is provided by the National Institute of Standards and Technology (NIST) Artificial Intelligence Risk Management Framework (AI RMF) 1.0 and the profile for GenAI [9,10]. For applications in quality management, this implies the need to design auditable solutions, ensure

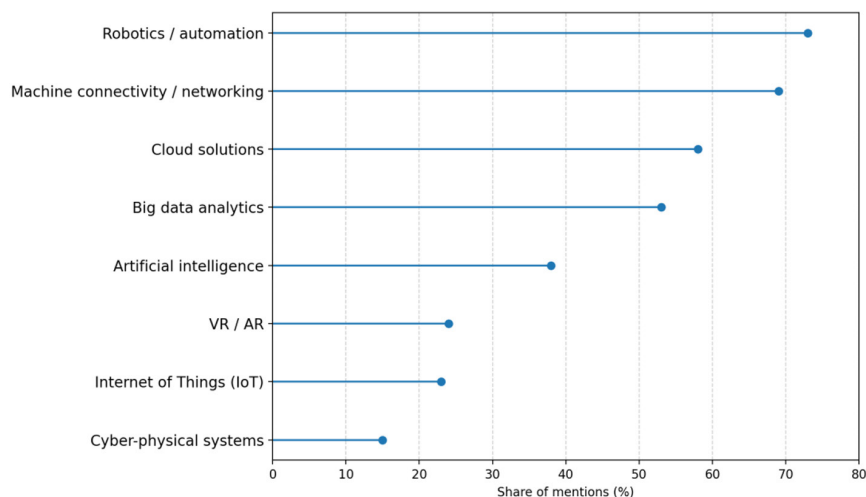


Figure 1. Relevance of Industry 4.0 technologies in 2024 [1]

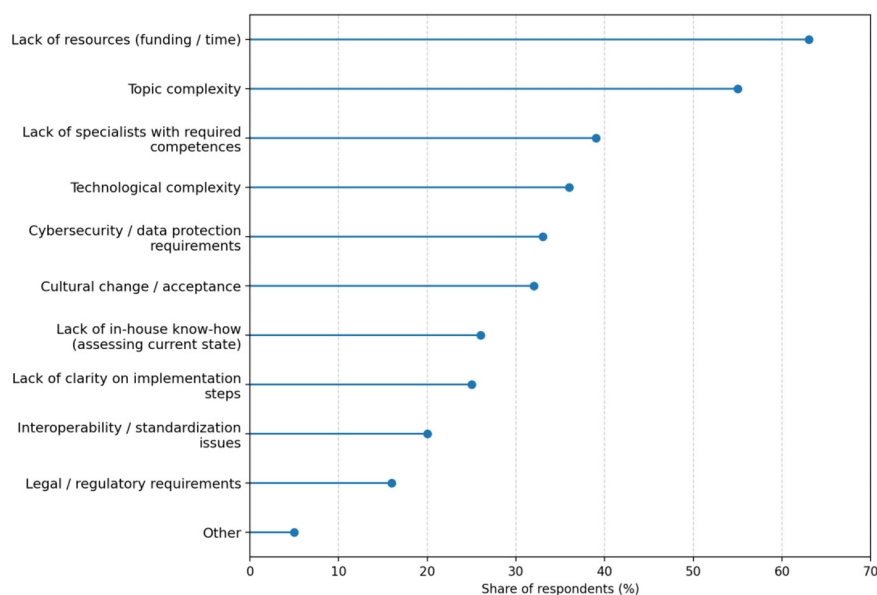


Figure 2. Key challenges in implementing Industry 4.0 components [2]

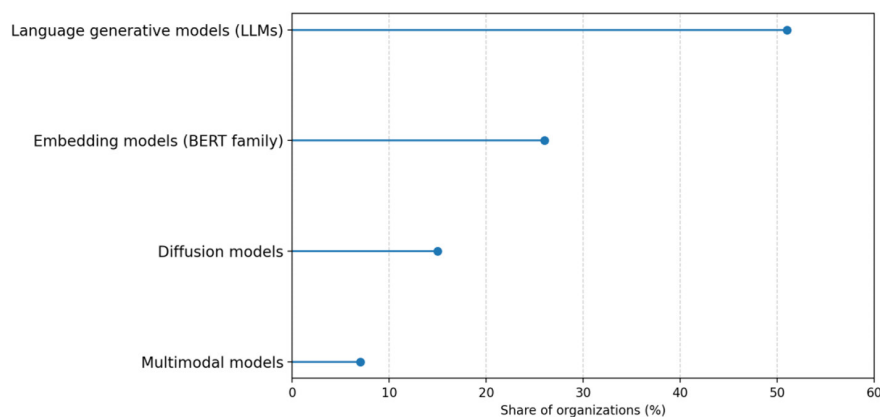


Figure 3. Planned use of model types in commercial deployments (2024) [3]

traceability of decisions, and include mechanisms to limit hallucinations and erroneous actions.

This paper focuses on computational methods and their applications in engineering, including artificial intelligence, production, and quality engineering. The research problem is formulated as the following question: how to design and evaluate a GenAI architecture for quality decisions that simultaneously (i) improves diagnostic efficiency and quality KPIs, (ii) maintains human accountability in the decision-making loop, and (iii) meets audit and AI governance requirements? In response, the paper proposes an LLM-augmented digital quality twin (LDQT) architecture and verifies its potential impact on quality metrics in a controlled simulation experiment. Given the interdisciplinary nature of this problem and the rapid change in the field of manufacturing, where

only production-ready, tested, and not experimental solutions are used, this verification is explicitly simulation-based and should not be read as evidence of completed industrial implementation. It provides insight into the strategy that is possible for future investigation in R&D departments, as the theoretical part is conceptually and statistically proven.

THEORETICAL FRAMEWORK

A digital twin is commonly defined as a formal, bidirectionally coupled representation of a physical object or process, encompassing data, models, and analytical services. The literature emphasises both the conceptual aspect (consistency of representation) and the operational

aspect (integration with the manufacturing lifecycle) [11,12]. Standardisation in manufacturing is developed, among others, in the ISO 23247 series, which organises concepts, modelling levels, and integration principles of digital twins in industrial automation [13,14]. In this work, the Digital Quality Twin is defined as a defect-centric, time-updated twin instance whose main objective is to represent critical-to-quality (CTQ) variables, conformity and non-conformity states, defect-mode probabilities, inspection states, rework/scrap flows, process-drift states, intervention constraints, and the expected consequences of corrective actions in a production flow. It differs from a generic manufacturing twin because its primary state variables are quality, risk, and decision safety rather than asset geometry or throughput alone. Incorporating this layer is consistent with the process approach and risk-based thinking in quality management [15].

LLMs are characterised by their ability to generate coherent narratives and syntheses from input data, which makes them attractive in settings where knowledge is scattered across documents and expert experience. However, their usefulness in high-risk environments depends on mechanisms that limit hallucinations. By attaching retrieved document content to the generation context, RAG constitutes the dominant family of enterprise-grade solutions [4]. In manufacturing, RAG implementations dedicated to quality control already appear, where the retrieval layer is integrated with repositories of process knowledge and engineering documentation [5]. The proposed approach extends this logic by a digital-twin component that tests recommendations in a what-if mode before they are implemented in the physical world; therefore, the LLM is treated as a diagnostic orchestrator rather than an autonomous controller.

There is a growing consensus that a key trajectory of AI deployments is a shift from task automation toward human-AI collaboration, in which AI systems operate as team members performing part of the cognitive work. This phenomenon is studied under the notions of AI-teaming and human-AI teaming, emphasising roles, responsibility, information exchange, and trust [16,17]. In the quality context, this means that LLMs should support hypothesis generation, prioritisation of tests, and communication of results (reporting), while decisions about process interventions remain approved by an authorised engineer.

Recent advances in production systems point to a renaissance of the agent-based approach, where LLMs act as planning and coordination components (agentic AI). From the perspective of digital twins, researchers emphasise the potential of combining multi-agent environments with LLMs to improve autonomy and flexibility in manufacturing systems [18]. A similar trend concerns advanced planning and scheduling (APS), for which architectures supported by LLMs are proposed, capable of interactively justifying planning decisions and adapting to disruptions [19]. This paper adapts that idea to the micro-scale of quality decisions: the LLM agent does not plan the entire schedule, but orchestrates the diagnostic process and recommends interventions, grounding them in knowledge and validating them in the twin.

Embedding GenAI in manufacturing processes requires the formalisation of AI governance. In practice, this means managing risk at the level of data, models, interfaces, and decisions. The NIST AI RMF 1.0 proposes a risk management framework built on the functions Govern–Map–Measure–Manage, and the profile for GenAI specifies risk categories characteristic of content generation, including hallucinations, data leakage, and vulnerability to context manipulation [9,10]. ISO/IEC 42001, in turn, provides system requirements for organisations (an AI management system) [8]. Finally, the AI Act establishes a baseline of regulatory expectations in the EU, which in industrial practice strengthens the need to design auditable systems with clearly defined responsibility and documentation [7].

An important stream of work at the intersection of automation, collaborative robotics, and AI methods is currently being developed intensively. For example, the usefulness of fuzzy logic for controlling cobots has been demonstrated [20], as well as the influence of process modelling using Business Process Model and Notation (BPMN) on collaborative robot deployments in enterprises [21]. The present study constitutes a logical extension of this stream. The paper analyses the GenAI layer as an interface and orchestrator of process knowledge, coupled with a digital quality twin and embedded in an AI governance regime.

Recent manufacturing literature also shows that performance evaluation is frequently grounded in parameterised process metrics and controlled experiments. Micro-milling studies emphasise the sensitivity of outcomes to tool geometry, tool treatment, and process-parameter choices

[22,23]. Efficiency metrics, neural-network mapping, and parameter-analysis approaches further illustrate how engineering studies operationalise performance assessment [24,25], while recent experimental and finite-element studies demonstrate the role of simulation in understanding process responses and deformation mechanisms [26,27]. Beyond process optimisation, AI adoption research highlights that usefulness, trust, and organisational conditions affect technology uptake [28]. These studies do not validate LDQT directly. However, they motivate treating the present work as a simulation-based, parameter-sensitive contribution rather than an industrial effectiveness claim. Industrial process-monitoring research based on machine-learning reconstruction for electrical tomography also demonstrates that non-invasive sensor data can be converted into diagnostic information for industrial processes [29]. At the maintenance-management level, AI-based decision trees and rough-set approaches have been used to support the assessment of lean maintenance and the formulation of production-maintenance decisions in manufacturing enterprises [30].

RESEARCH PROBLEM, OBJECTIVE AND HYPOTHESES

Based on the above framework, a literature-method gap is identified. Despite the growing number of RAG/LLM prototypes for industry, there is a lack of quantitative and, therefore, comparable computational evidence regarding (a) the extent to which integrating LLMs into quality decision-making improves quality KPIs such as first-pass yield (FPY), scrap, and cost of quality, and (b) how strongly the outcome depends on governance mechanisms limiting erroneous interventions. In particular, the literature rarely combines within one model: (i) RAG as a grounding layer, (ii) a digital twin as a tool for recommendation validation, and (iii) human-in-the-loop as an accountability mechanism.

The novelty of LDQT is therefore not the isolated use of LLMs, RAG, or digital twins, but their operational coupling into an auditable quality-decision loop: (i) RAG provides cited evidence, (ii) the Digital Quality Twin simulates the intervention and checks operational constraints, (iii) the governance layer records and gates the action, and (iv) the human approver retains

accountability. The proposed artefact is evaluated at the metric level rather than only described conceptually.

The aim of this paper is to design an auditable LDQT architecture and verify, in a simulation experiment, its potential impact on the effectiveness and safety of quality decisions. The following research hypotheses are formulated:

- H1: In the simulation model, the LDQT architecture shortens the response time to process drift (proxy: detection delay) compared with the manual and rule-based scenarios.
- H2: In the simulation model, LDQT improves quality metrics (FPY, scrap) and reduces the unit cost of quality + downtime relative to the manual scenario.
- H3: In the simulation model, the governance layer (RAG + twin-based validation + human-in-the-loop) reduces the risk of erroneous interventions without losing efficiency gains.

METHODOLOGY

Methodologically, the paper combines the design science approach (designing the LDQT architecture as an artefact) with a computational experiment in which we investigate the effects of different intervention policies in a production stream. The research workflow (from gap identification to statistical analysis) is shown in Figure 4. The study is intentionally limited to simulation and does not claim validation through a pilot implementation, an industrial dataset, or a live factory prototype.

Proposed LDQT architecture

The LDQT architecture (Figure 5) comprises five functional layers: (1) acquisition and standardisation of process data from the Industrial Internet of Things (IIoT), manufacturing execution systems (MES), supervisory control and data acquisition (SCADA) systems, and quality artifacts, (2) a knowledge base with vectorisation and retrieval mechanisms, (3) an RAG layer ensuring evidence grounding, (4) an LLM agent generating diagnostic hypotheses and recommendations, and (5) a digital quality twin used to validate what-if scenarios. The architecture is integrated by an AI governance layer (policies, traceability, access control, model registry, approval workflow, and audit), while the human

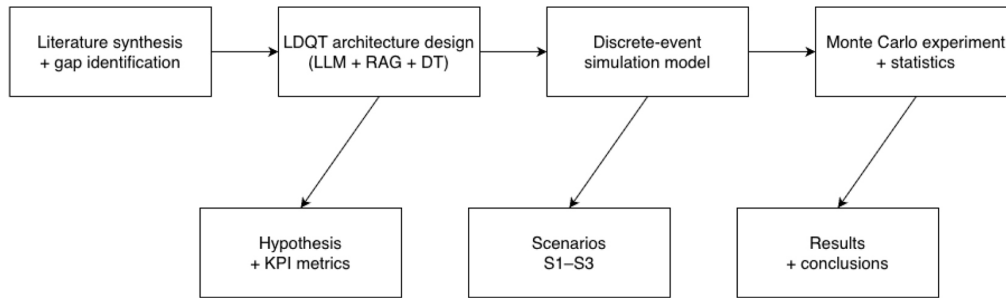


Figure 4. Methodological workflow of the study: from literature synthesis to the Monte Carlo experiment

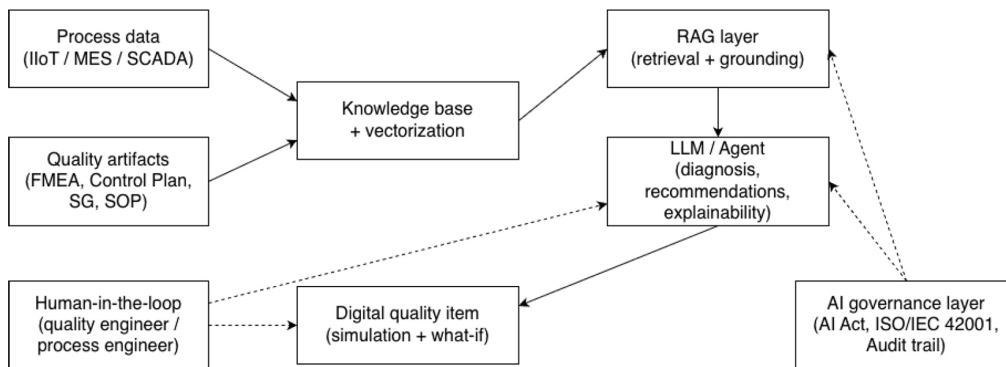


Figure 5. Conceptual architecture of the LLM-augmented digital quality twin (LDQT)

remains in the decision loop. In practice, this means that the LLM is not a source of truth but an orchestrator: it combines signals, documents, and simulation results into a coherent recommendation that can be justified and verified.

Governance layer and twin-based recommendation verification

The governance layer treats each LLM output as a proposal, not an executable command. The proposed sequence is: (1) an anomaly or drift signal is detected; (2) the RAG module retrieves source fragments from approved repositories such as the control plan, FMEA, 8D report, SOP, and maintenance log; (3) the LLM forms a recommendation packet containing evidence identifiers, assumptions, proposed action, expected quality effect, and confidence/risk score; (4) the Digital Quality Twin runs a what-if simulation using the current process state; (5) the packet is routed to an authorised quality engineer or production supervisor for approval; and (6) only the approved action is transferred to MES/SCADA or standard work instructions.

To prevent wrong AI recommendations, the gate rejects or escalates outputs with missing evidence, contradictory retrieved sources, low

retrieval score, infeasible parameter limits, negative what-if effect, cybersecurity or data-quality flags, or lack of human approval. High-risk interventions require dual approval by the quality engineer and process owner, plus a rollback and monitoring plan. The LLM has no direct write access to programmable logic controllers (PLCs), MES, or SCADA.

The audit log records the event identifier, timestamp, process state, data-source versions, retrieved document identifiers and hashes, prompt template, model version, recommendation text, twin input and output, risk score, approver identity and role, final decision, execution status, and post-intervention outcome. These records support traceability, root-cause analysis, and periodic review of model drift.

Operationalisation of hypotheses and metrics

To enable an unambiguous verification of the hypotheses, the constructs were mapped to metrics and measurement procedures. The mapping is summarised in Table 1. We adopt measures typical for quality engineering, including first-pass yield (FPY) and scrap, complemented by

decision-safety indicators (probability of incorrect intervention) and a unit cost metric aggregating the effects of defects, rework, and downtime. All economic values are expressed in arbitrary units (a.u.) to enable comparison of scenarios without disclosing plant-specific cost data.

Simulation model and experimental scenarios

The empirical layer of the study takes the form of a discrete-event simulation (DES) at the level of a production stream. The model describes the production of N items, in which stochastic process-drift events occur (e.g., tool miscalibration, change in material properties). Drift increases the defect probability until it is detected and a corrective intervention is performed. The intervention decision depends on the scenario: S1 (manual) represents classic analysis by a quality engineer, S2 (rule-based) represents a decision supported by a simple rule-based decision-support system (DSS), and S3 (LDQT) represents a decision supported by an LLM-RAG agent with twin-based validation and a human-in-the-loop mechanism. Scenario parameters are listed in Table 2. They are transparent, engineering-plausible assumptions selected to represent three levels of diagnostic maturity rather than empirically calibrated parameters of a specific plant.

The values in Table 2 are intended to be realistic at the level of order of magnitude, not universal benchmarks. With a mean cycle time of approximately two minutes, the manual delay of 220 items represents batch- or shift-level detection, the rule-based delay of 140 items represents partial automation of alarm interpretation, and the LDQT delay of 80 items represents earlier diagnosis supported by retrieval, hypothesis generation, and what-if testing. The defect probabilities describe a stable process with a low baseline non-conformity rate (1%), a drift state with a sixfold

higher defect risk (6%), and an incorrect intervention state with additional degradation (8%). The assumed incorrect-intervention probabilities (5%, 3%, and 2%) express the expected effect of increasing decision support and governance. Because these assumptions remain uncertain, the Results section includes a sensitivity analysis.

The experiment was conducted using a Monte Carlo simulation: for each scenario, $R=400$ replications were executed, and each replication comprised $N=4000$ items. Metric distributions were evaluated, 95% confidence intervals for means were computed, and effect sizes (Cohen's d) with bootstrap uncertainty estimation were calculated. For H1 and H2, scenario means from independent Monte Carlo replications were compared using two-sided Welch's t -tests, with degrees of freedom estimated using the Welch-Satterthwaite approximation. For H3, the incorrect-intervention share was compared using a two-proportion z -test based on aggregated intervention-event counts, not item counts. The corresponding test statistics and, where applicable, degrees of freedom are reported in Table 4. This procedure makes it possible not only to identify differences in means but also to assess their practical importance (effect magnitude) under process variability.

RESULTS

Aggregated experiment results are presented in Table 3. Under the adopted simulation assumptions, a systematic improvement of all quality KPIs is visible in the LDQT scenario: first-pass yield increases, the scrap fraction decreases, and the unit cost of quality + downtime is reduced. These values describe the model output, not verified factory performance.

Table 4 clarifies that H1 and H2 were tested through two-sided Welch's t -tests on independent

Table 1. Operationalisation of research constructs and metrics

Hypothesis	Construct	Metric	Operationalisation in the study
H1	Diagnostic efficiency (mean time to root-cause analysis, MTRCA)	Mean drift detection delay [items]	$\mu(\text{det_delay}) - \text{model parameter and number of items in drift state}$
H2	Quality effectiveness	First-Pass Yield (FPY)	$1 - \text{number of defects} / N$
H2	Quality effectiveness	Scrap rate	$\text{number of scrap items} / N$
H2	Economic efficiency	Unit cost of quality + downtime	$(C_{\text{scrap}} \cdot \text{scrap} + C_{\text{rework}} \cdot \text{rework} + C_{\text{dt}} \cdot \text{downtime}) / N$
H3	Decision safety	Probability of incorrect intervention	$\text{wrong_interventions} / \text{interventions}$

Table 2. Key parameters of the three scenarios in the simulation experiment

Scenario	Description	μ detection delay [items]	σ detection delay [items]	P(incorrect intervention)	P(defect normal)	P(defect drift)	P(defect after incorrect)
S1	Manual (baseline)	220	60	0.05	0.01	0.06	0.08
S2	Rule-based DSS	140	45	0.03	0.01	0.06	0.08
S3	LLM-RAG + governance (Retrieval-Augmented Generation + twin validation + human approval)	80	35	0.02	0.01	0.06	0.08

Table 3. Monte Carlo experiment results: mean and 95% CI for key metrics

Scenario	FPY	Scrap	Cycle [min]	Cost [a.u./item]	Interventions
S1: Manual (baseline)	0.9836 (0.9832–0.9841)	0.0066 (0.0064–0.0068)	2.0692 (2.0669–2.0715)	1.3118 (1.2724–1.3512)	2.21
S2: Rule-based DSS	0.9855 (0.9851–0.9858)	0.0059 (0.0057–0.0061)	2.0632 (2.0613–2.0650)	1.1772 (1.1458–1.2086)	2.46
S3: LLM-RAG + governance	0.9873 (0.9870–0.9876)	0.0052 (0.0051–0.0054)	2.0573 (2.0556–2.0589)	1.0453 (1.0200–1.0707)	2.58

Table 4. Explicit statistical verification of hypotheses in the simulation study

Hypothesis/comparison	Main metric	Difference or effect	Test and result	Interpretation
H1: S3 vs S1	Detection delay	80 vs 220 items; Cohen's d = -2.85	Welch's t-test: t = -40.31; df = 642.36; p < 0.001	Supported in simulation
H1: S3 vs S2	Detection delay	80 vs 140 items; Cohen's d = -1.49	Welch's t-test: t = -21.05; df = 752.41; p < 0.001	Supported in simulation
H2: S3 vs S1	Cost	-0.2665 a.u./item (-20.3%); d = -0.79	Welch's t-test: t = -11.18; df = 681.02; p < 0.001	Supported for cost
H2: S3 vs S1	FPY	+0.0037; d = 0.95	Welch's t-test: t = 13.45; df = 695.16; p < 0.001	Supported for yield
H3: S3 vs S1	Incorrect intervention probability	0.02 vs 0.05	Two-proportion z-test: z = -3.52; p < 0.001; df n/a	Supported as a scenario-based governance result

Monte Carlo replications, with the Welch-Satterthwaite approximation used for the degrees of freedom. H3 is interpreted as a governance-design comparison because the incorrect-intervention probability is an explicit scenario parameter; its pairwise comparison was reported using a two-proportion z-test. The statistical results, therefore, support the internal validity of the simulation experiment but do not replace factory validation.

Differences in unit costs are not only reflected in mean values, but span the entire distribution. Figure 6 presents the distribution (boxplot) of cost per item; in the LDQT scenario, we observe a downward shift of the distribution and a smaller number of extreme observations. In interpretive terms, LDQT reduces not only the typical simulated cost of quality but also the risk of high-cost episodes resulting from prolonged drift or incorrect interventions.

Because quality decisions are dynamic in nature, we also analysed defect evolution over

time. Figure 7 shows the cumulative defect rate in a single simulation run. The LDQT curve stabilises at a lower level than in the manual scenario, which is consistent with hypothesis H1 within the simulated environment: faster drift detection reduces the number of defective items produced before corrective action.

To frame the results in the language of applied statistics, effect sizes (Cohen's d) were computed relative to the baseline scenario for major metrics (cost, scrap, cycle time, FPY). Figure 8 shows that the improvement in cost and scrap in the LDQT scenario is at least medium to large (in terms of |d|), with narrow uncertainty intervals. This finding strengthens H2 in the simulation study and suggests that the benefits are not an artefact of individual runs but a stable outcome of the faster and safer intervention mechanism.

To address the dependence on assumed parameters, a one-factor-at-a-time sensitivity

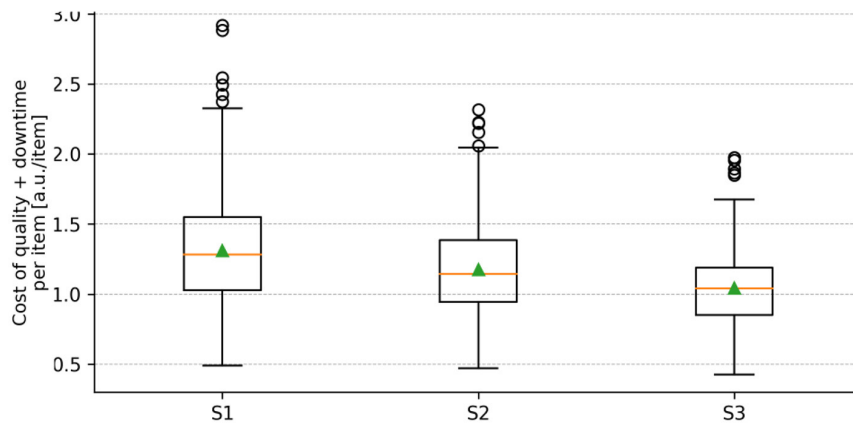


Figure 6. Distribution of the unit cost of quality + downtime in three scenarios

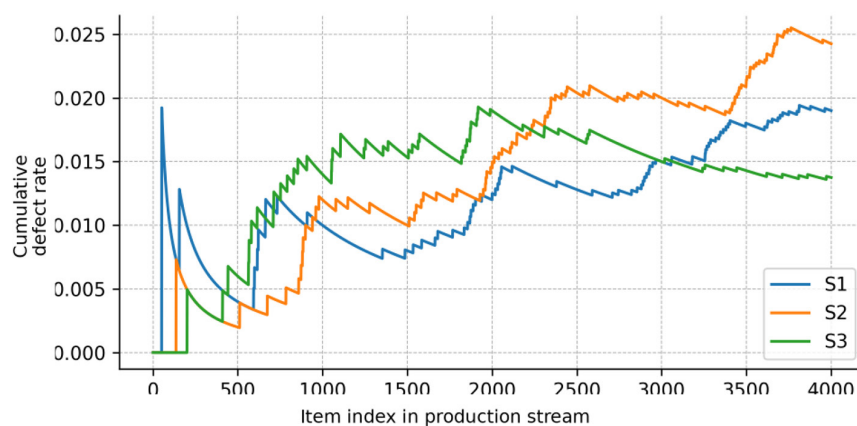


Figure 7. Quality dynamics: cumulative defect rate over time

analysis was added. The analysis varied LDQT detection delay, incorrect-intervention probability, and drift severity while keeping the comparison structure unchanged. The results in Table 5 show that the direction of the LDQT advantage

remains stable, although the magnitude decreases under pessimistic assumptions.

The sensitivity analysis indicates that the conclusions are robust in direction but not in exact magnitude. This supports a cautious interpretation: LDQT is promising in the model, yet

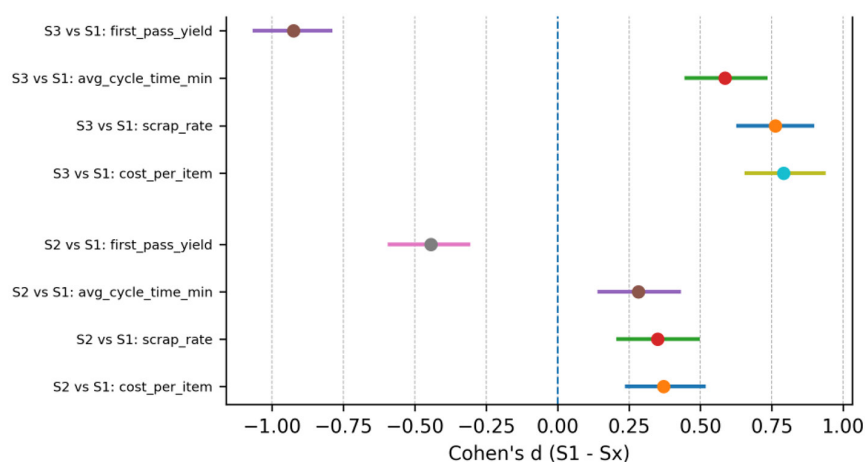


Figure 8. Effect sizes (Cohen's d) relative to the baseline scenario (bootstrap 95% CI; own elaboration)

Table 5. Sensitivity analysis of LDQT advantage over the manual baseline

Variant	Parameter change	LDQT cost advantage vs S1	FPY direction	Interpretation
Base model	Table 2 values	20.3% lower cost	Higher	Reference case
Pessimistic detection	S3 mean delay +25% (80 -> 100 items)	17.1% lower cost	Higher	Benefit remains
Weaker governance	S3 incorrect-intervention probability doubled (0.02 -> 0.04)	18.5% lower cost	Higher	Robust but smaller
More severe drift	P(defect drift) +25% (0.06 -> 0.075)	19.4% lower cost	Higher	Faster response still limits exposure
Boundary case	S3 mean delay +50% and P(incorrect) = 0.04	14.8% lower cost	Higher, reduced	Main direction persists

plant-specific calibration is necessary before operational claims can be made.

DISCUSSION

The simulation results should be interpreted as computational evidence that a certain decision architecture is consistent with the expected direction of KPI improvement, rather than as a direct measure of the effectiveness of any specific LLM or as proof of industrial readiness. In the LDQT model, intelligence is represented through parameters that shorten drift-detection time and reduce the risk of incorrect interventions, corresponding to an organisational effect: faster hypothesis generation, better use of knowledge, and validation of recommendations in a controlled twin environment. The conceptual difference between LDQT and adjacent approaches is summarised in Table 6.

This comparison highlights the main contribution of the paper that LDQT integrates evidence grounding, digital-twin validation, human accountability, and auditable governance in one quality-decision loop.

First, the observed effect is consistent with the growing stream of research on RAG for quality control. Studies on advanced RAG systems for manufacturing quality indicate that grounding in documents and data increases the usefulness of GenAI for engineers [5]. The present study complements that stream with a KPI perspective by showing, under explicit assumptions, how shortening the time from signal to action can translate into a reduction of the cost of quality and the risk of high-cost episodes.

Second, the incorporation of a digital quality twin implements recommendations stemming from the concept of manufacturing digital twins, i.e., treating the twin as an environment integrating data and models across the production life-cycle [11,13]. In practice, LDQT interprets the twin as a pre-verification tool for interventions (what-if), which directly matters for minimising erroneous actions. This is particularly important in systems with high technological complexity (cf. Figure 2), where the consequences of incorrect settings may escalate cascadingly.

Third, the results align with findings from human-AI teaming research, which indicate that team productivity depends on a clear allocation of roles, trust, and the quality of feedback

Table 6. Conceptual comparison of LDQT with related manufacturing decision-support approaches

Approach	Main strength	Typical gap	LDQT extension
Standalone RAG quality assistant	Evidence-grounded answers from documents	No process-state validation before action	Adds Digital Quality Twin, what-if checking, and approval gate
Classical digital quality twin	Process simulation and quality-state representation	Limited natural-language synthesis of unstructured knowledge	Adds LLM-RAG interface and an auditable recommendation packet
Rule-based DSS	Transparent execution of predefined rules	Brittle for new defect patterns and unstructured documentation	Adds retrieval, explanation, and simulation-based recommendation screening
Generic AI copilot	Fast narrative support for engineers	Weak traceability and execution control	Adds source records, model/twin logs, human sign-off, and no direct write access

loops [16,17]. LDQT operationalises these postulates through the human-in-the-loop principle and traceability. The LLM recommendation is linked to evidence (RAG), a twin-based test, and a named approver. As a consequence, the system augments the competence of the quality engineer rather than shifting responsibility to the model.

Fourth, the agentic-AI and multi-agent digital-twin trend suggests that the future of manufacturing systems will rely on the coordination of multiple decision components, with LLMs as an interaction and planning layer [18,19]. LDQT can be interpreted as a specialised case of such a system: an LLM agent manages the diagnostic workflow and integrates results, yet its actions are constrained by governance and validation.

Finally, compliance and AI governance are an essential part of the discussion. The LDQT scenario assumes mechanisms aligned with practices recommended by ISO/IEC 42001 and NIST (risk management, measurement, continuous improvement) [8,9]. At the design level, this entails separating the generative layer from the execution layer, controlling access to data, maintaining an audit log, and establishing procedures for impact assessment. This approach is also consistent with the EU regulatory direction [7].

Practical deployment would require integration with MES/SCADA, quality-management repositories, and access-control infrastructure. The computational layer would include a vector database, an LLM endpoint or local model, a simulation service for the Digital Quality Twin, monitoring dashboards, and secure interfaces to existing IT/OT systems. Main cost drivers include data cleaning, document governance, model validation, cybersecurity hardening, staff training, and maintenance of prompts, embeddings, model versions, and twin parameters. From an adoption perspective, technical accuracy alone is insufficient, usefulness, trust, explainability, organisational support, and workflow fit are likely to influence whether engineers accept the system [28].

LIMITATIONS

The study has limitations typical of computational experiments and should be interpreted accordingly. First, the simulation model is stylised: it represents drift and interventions at a probabilistic level rather than through full process physics. Second, LLM behaviour is represented by

parameters (detection time, error risk) rather than by executing a specific model on real data. Third, costs and times are relative and require calibration for a particular plant. Fourth, no pilot implementation, industrial dataset, prototype demonstration on live factory logs, or closed-loop MES/SCADA deployment was performed; therefore, the results cannot confirm practical effectiveness in a real manufacturing setting. Fifth, the selected parameters were justified as engineering-plausible assumptions and explored through sensitivity analysis, but they are not universally transferable benchmarks across plants. Additional limitations concern cybersecurity risks (prompt injection, retrieval poisoning, data leakage, unauthorised access), data-quality problems in source repositories, model drift, long-term maintenance of vector databases and prompts, and the operational cost of integration, validation, monitoring, and staff training. Nevertheless, this construction is purposeful: it isolates mechanisms and allows hypotheses to be tested in a controlled environment.

CONCLUSIONS

This paper proposed an auditable LLM-augmented Digital Quality Twin architecture that combines GenAI (LLM) with RAG, a digital quality twin, and the human-in-the-loop principle for production and technology management. The computational experiment indicates that, under the adopted assumptions, LDQT may shorten drift-detection delay, improve first-pass yield, reduce scrap, and lower the unit cost of quality and downtime relative to manual and rule-based scenarios. The explicit governance layer is central because it converts LLM output into an auditable recommendation packet that must be grounded in retrieved evidence, checked in the digital quality twin, and approved by a responsible human. The results should not be interpreted as proof of real-factory effectiveness. They show that the architecture is internally consistent and promising in simulation; practical effectiveness can be confirmed only after calibration with industrial data and validation in a production environment. The main contribution is a defined Digital Quality Twin concept, an auditable LDQT architecture, and a simulation-based metric set for balancing efficiency and decision safety in Industry 4.0/5.0 quality management.

REFERENCES

1. BearingPoint. Industrie 4.0 - Relevanz technologischer Entwicklungen der Industrie 4.0 in 2024. Statista; 2025.
2. BearingPoint. Industrie 4.0 - Herausforderungen durch Industrie 4.0 in 2024. Statista; 2025.
3. AI Infrastructure Alliance. Firms planned LLM model usage in commercial deployments worldwide 2024. Statista; 2024.
4. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv Neural Inf Process Syst*. 2020;33:9459–74. <https://doi.org/10.48550/arXiv.2005.11401>
5. Heredia Álvaro JA, González Barreda J. An advanced retrieval-augmented generation system for manufacturing quality control. *Adv Eng Inform*. 2025;64:103007. <https://doi.org/10.1016/j.aei.2024.103007>
6. Breque M, De Nul L, Petridis A. Industry 5.0: Toward a sustainable, human-centric and resilient European industry. Luxembourg: Publications Office of the European Union; 2021.
7. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [Internet]. Official Journal of the European Union; 2024. Available from: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
8. International Organization for Standardization. ISO/IEC 42001:2023 Information technology - Artificial intelligence - Management system [Internet]. Geneva: ISO; 2023. Available from: <https://www.iso.org/standard/81230.html>
9. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg: NIST; 2023.
10. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1) [Internet]. Gaithersburg: NIST; 2024. Available from: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf>
11. Villegas LF, Macchi M, Polenghi A. Digital twins in manufacturing: A unified conceptual framework. *Annu Rev Control*. 2025;60:101031. <https://doi.org/10.1016/j.arcontrol.2025.101031>
12. Shao G, Helu M. Framework for a digital twin in manufacturing: Scope and requirements. *Manuf Lett*. 2020;24:105–7. <https://doi.org/10.1016/j.mfglet.2020.04.004>
13. International Organization for Standardization. ISO 23247-1:2021 Automation systems and integration - Digital twin framework for manufacturing - Part 1: Overview and general principles [Internet]. Geneva: ISO; 2021. Available from: <https://www.iso.org/standard/75066.html>
14. Shao G, Frechette S, Srinivasan V. An analysis of the new ISO 23247 series of standards on digital twin framework for manufacturing. In: Proceedings of the ASME 2023 18th International Manufacturing Science and Engineering Conference, MSEC 2023. Gaithersburg: National Institute of Standards and Technology; 2023. Available from: <https://www.nist.gov/publications/analysis-new-iso-23247-series-standards-digital-twin-framework-manufacturing>
15. International Organization for Standardization. ISO 9001:2015 Quality management systems - Requirements [Internet]. Geneva: ISO; 2015. Available from: <https://www.iso.org/standard/62085.html>
16. Schmutz JB, Outland N, Kerstan S, Georganta E, Ulfert AS. AI-teaming: Redefining collaboration in the digital era. *Curr Opin Psychol*. 2024;58:101837. <https://doi.org/10.1016/j.copsyc.2024.101837>
17. Berretta S, Tausch A, Ontrup G, Gilles B, Peifer C, Kluge A. Defining human-AI teaming the human-centered way: A scoping review and network analysis. *Front Artif Intell*. 2023;6:1250725. <https://doi.org/10.3389/frai.2023.1250725>
18. Greis NP, Cherukuri HP, Outeiro JCM. Multi-agent systems for manufacturing digital twins: A perspective on agency and large language models. *IFAC-PapersOnLine*. 2025;59(10):1612–7. <https://doi.org/10.1016/j.ifacol.2025.09.271>
19. Li M, Zhou Q, Li W, Qu T, Yang M, Jiang P. A4PS: Agentic AI-assisted advanced planning and scheduling with large language models for smart manufacturing. *J Manuf Syst*. 2026;85:207–26. <https://doi.org/10.1016/j.jmsy.2026.01.003>
20. Cioch M, Kulisz M, Kański Ł. Implementing AI collaborative robots in manufacturing - Modeling enterprise challenges in Industry 5.0 with fuzzy logic. *Adv Sci Technol Res J*. 2024;18(7):229–38. <https://doi.org/10.12913/22998624/192833>
21. Pizoń J, Kański Ł, Kulisz M, Chadam J, Sánchez García E. Using Business Process Model and Notation 2.0 to deploy cobots in a manufacturing system - Case study. *Adv Sci Technol Res J*. 2024;18(7):150–62. <https://doi.org/10.12913/22998624/192491>
22. Rathod VP, Wankhade SH. A review on impact of micro-tools on micro-milling outcomes for aluminium alloy. *Advances in Science and Technology Research Journal*. 2025;19(3):322–340. <https://doi.org/10.12913/22998624/200007>
23. Rathod VP, Wankhade SH. Comprehensive review of tool treatments and innovations in micro-milling precision and performance. *Advances in Science and Technology Research Journal*. 2025;19(4):241–257. <https://doi.org/10.12913/22998624/200545>
24. Rathod VP, Wankhade SH, Kamble YG. Evaluating

- the efficiency of micro end mill tools and performance metrics in micro milling. *Journal of Mines, Metals and Fuels*. 2025;73(7):2253–2265. <https://doi.org/10.18311/jmmf/2025/49414>
25. Kamble YG, Rathod VP, Pantawane P, Wankhade SH, et al. Center drill geometry mapping in friction drilling of aluminium 6063 using artificial neural network approach. *Advances in Science and Technology Research Journal*. 2025;19(11):124–148. <https://doi.org/10.12913/22998624/209120>
26. Rathod VP, Wankhade SH, Kamble YG. Experimental analysis of process parameters in micro-milling of AA6063-T6 using 7% cobalt WC tools. *Advances in Science and Technology Research Journal*. 2026;20(1):162–182. <https://doi.org/10.12913/22998624/210942>
27. Kamble YG, Rathod VP, Pantawane P, Wankhade SH, et al. Cutting tool and workpiece deformation in friction drilling of aluminum alloy 6063 by finite element model simulation. *Advances in Science and Technology Research Journal*. 2026;20(6):343–355. <https://doi.org/10.12913/22998624/218094>
28. Fathima GN, Rathod VP, Mishra P, et al. Improving the adoption of artificial intelligence in library financial services: A comprehensive UTAUT2 framework with AI-specific variables. *Library of Progress: Library Science, Information Technology and Computer*. 2024;44(3).
29. Rymarczyk T, Kłosowski G, Kozłowski E, Tchórzewski P. Comparison of selected machine learning algorithms for industrial electrical tomography. *Sensors*. 2019;19(7):1521. <https://doi.org/10.3390/s19071521>
30. Antosz K, Pasko L, Gola A. The use of artificial intelligence methods to assess the effectiveness of lean maintenance concept implementation in manufacturing enterprises. *Appl Sci*. 2020;10(21):7922. <https://doi.org/10.3390/app10217922>