Human Factors in Industrial and Systems Engineering, Vol. 219, 2026, 138–145
https://doi.org/10.54941/ahfe1007784

Design and Implementation of a Knowledge-Based Assistance System for Smart Failure Management in Manufacturing SMEs Using Large Language Models

Turgut Refik Caglar

Technical University of Berlin, Quality Science, Pascal Str. 8-9, 10587 Berlin, Germany

ABSTRACT

Manufacturing small and medium-sized enterprises (SMEs) are under increasing pressure to detect, analyse, and eliminate quality-related failures faster and more systematically, while operating with limited personnel, fragmented information structures, and heterogeneous IT environments. Although standards such as ISO 9001 require structured corrective action and documented organizational learning, practical failure management in many SMEs remains reactive, media-discontinuous, and weakly connected to reusable knowledge. This paper presents the design and prototypical implementation of a modular, knowledge-based assistance system that combines established quality engineering methods with natural language processing (NLP), machine learning, and large language models (LLMs) to support the entire problem-solving cycle. The research follows a design science research approach. First, requirements were derived from normative sources, the state of research, and an empirical industry survey with 104 valid company responses. The resulting requirements were structured into seven functional domains. On this basis, a modular reference architecture was developed that integrates structured failure capture, historical document analysis, method-guided problem solving, an explainable knowledge base, and feedback-driven learning loops. The prototype was implemented as a containerized full-stack web application using open-source technologies, including Flask, PostgreSQL, Docker, HTML/CSS/Bootstrap/JavaScript, and Llama 3- and Rasa-based conversational services. Transformer-based subcomponents for root-cause classification and guided 5-Why questioning complement the LLM-supported retrieval-augmented generation (RAG) assistant. The system was prototypically validated using historical failure documentation from manufacturing case studies. Evaluation results indicate improvements in knowledge accessibility, reduction in analysis time, increased consistency in root cause identification, and enhanced standardization of corrective action documentation. The findings suggest that AI-enhanced assistance systems can significantly strengthen organizational learning capabilities in SMEs, provided that they are embedded within structured quality management frameworks. The paper contributes to research in digital quality management by (1) providing a structured requirement-based reference architecture for AI-supported failure management systems, (2) proposing a systematic mapping between data mining methods and problem-solving phases, and (3) demonstrating a practical integration approach for LLMs in industrial quality environments. It bridges the gap between classical quality

Received March 20, 2026; Revised April 29, 2026; Accepted May 14, 2026; Available online July 20, 2026

engineering and modern AI-based knowledge systems, offering a scalable pathway toward intelligent, learning-oriented failure management in manufacturing.

Keywords: Smart failure management, Knowledge-based assistance system, Large language models (LLMs), Retrieval-augmented generation (RAG)

INTRODUCTION

Failures in manufacturing remain a major economic and organizational burden. They consume time, tie up qualified personnel, increase rework and scrap, and often interrupt the flow of production. At the same time, many companies still handle failure information through handwritten notes, spreadsheets, isolated reports, and department-specific databases. This fragmentation makes it difficult to identify patterns, reuse successful corrective actions, and institutionalize learning across cases and product generations.

The challenge is particularly pronounced in SMEs. Compared with large enterprises, SMEs often operate with fewer quality specialists, limited digital infrastructure, and lower methodological redundancy. Yet they are expected to manage increasingly complex products, stricter customer requirements, and shorter response times. Consequently, the operational handling of failures is frequently shaped by tacit expert knowledge and ad hoc communication rather than by a consistent, digitally supported problem-solving proces. (Caglar & Jochem, 2023).

Existing digital failure management solutions address only parts of this problem. Some emphasize documentation and workflow control; others focus on isolated analytics or machine learning applications. Few provide an integrated environment that combines structured quality methods, historical failure knowledge, semantic search, and interactive decision support. This paper addresses that gap by proposing a modular assistance system for smart failure management in manufacturing SMEs. The system is designed not to replace established quality engineering practice, but to operationalize and augment it through knowledge processing, explainable AI, and conversational user support.

RESEARCH METHOD AND REQUIREMENT DERIVATION

The study follows a design science research (DSR) logic in which a practically relevant artifact is developed on the basis of a rigorous problem analysis, iteratively specified, implemented, and evaluated. The starting point was a triangulated requirement analysis combining three sources: normative frameworks, especially ISO 9001; the state of research on failure management, knowledge-based systems, and data-driven quality management; and an empirical industry survey designed to validate and refine practical needs (Caglar et al., 2025a).

The empirical survey addressed manufacturing companies from several sectors and yielded 104 complete and usable responses. The results confirmed that many firms already collect failure data, but predominantly in Excel files, local databases, or paper-based forms. At the same time, respondents

emphasized the practical importance of user-friendly interfaces, structured method support, KPI dashboards, the import of historical failure data, and AI-based processing of free-text failure descriptions. The survey therefore did not merely validate literature-based assumptions; it also sharpened the implementation priorities for an SME-oriented assistance system.

From this analysis, 20 requirements were derived and grouped into seven functional domains (see Table 1) to create a Knowledge-Based Assistance System for Smart Failure Management (SFM). These domains cover both the operational core of failure handling and system-level properties required for sustainable deployment in industrial practice.

Table 1: Functional requirement domains of the proposed system.

Domain	Primary Objective	Representative Capabilities
Failure acquisition	Capture failures close to the point of occurrence	Digital input forms, contextual product/process metadata, structured event logging
Documentation	Create reusable and standardized failure knowledge	Consistent records, update logic, access rights, central knowledge base
Decision support mechanisms	Interpret incomplete or vague failure descriptions	Free-text processing, semantic retrieval, contextual search
Root cause and corrective action support	Guide users from symptoms to validated actions	Cause suggestions, action proposals, user feedback loops
Structured problem-solving guidance	Operationalize quality and analytics methods	PDCA-oriented flow, 8D/FMEA support, guided method selection
Implementation control	Track follow-up and effectiveness	KPI monitoring, reporting, status management, exportable documentation
System-level learning	Continuously improve knowledge quality and usability	Historical case ingestion, role-based maintenance, modular extensibility

METHODOLOGY

The resulting architecture follows a modular logic in order to support scalability, role-specific interaction, and adaptation to company-specific needs. Conceptually, the system is organized around the lifecycle of a failure event. A schematic representation of the modular structure is shown in Figure 1.

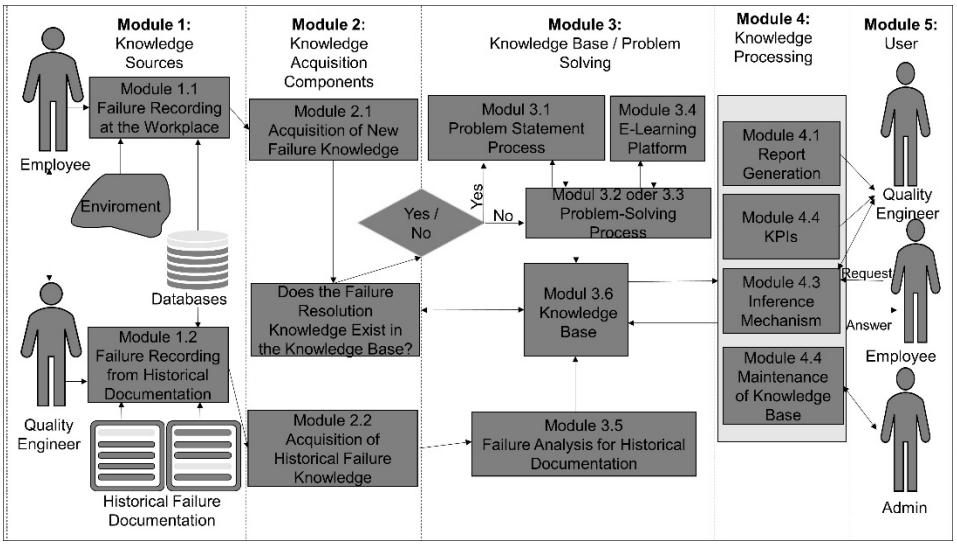

Figure 1: General procedural model for the development of the knowledge-based SFM.

Module 1 identifies and captures relevant knowledge sources. Current failure events are entered by shop-floor users through structured interfaces, while historical documentation such as 8D reports, FMEA sheets, and company-specific archives is collected by quality engineers for systematic reuse.

Module 2 transforms these heterogeneous inputs into machine-readable knowledge. This includes preprocessing, contextual enrichment, and classification. New cases are prepared for operational problem solving, while historical documents are segmented into reusable knowledge elements covering the core triad of failure description, cause, and corrective action.

Module 3 realizes the actual problem-solving and knowledge-base functions. If a new case matches an existing pattern, the user receives context-sensitive support based on comparable historical cases. If no sufficiently similar case exists, the system initiates a structured problem-solving pathway. Here, deterministic quality methods such as Ishikawa, 5-Why, FMEA, and 8D reporting are embedded into a broader problem-solving cycle and digitally supported through guided workflows. The structure of module 3 is discussed in other works from Caglar (Caglar et al., 2025b).

Module 4 is responsible for knowledge processing and system learning. It consolidates validated outcomes, updates the knowledge base, tracks KPIs, and generates structured reports. Module 5 defines role-specific access and interaction patterns for shop-floor employees, quality managers, and administrators. Together, these modules create a coherent assistance system that connects operational failure handling with organizational learning.

TECHNICAL IMPLEMENTATION

The prototype was implemented as a full-stack web application grounded in open-source technologies (see Table 2). This decision was driven by three practical requirements: cost efficiency, transparency, and adaptability. The

complete application was containerized with Docker in order to ensure reproducible deployment, isolated services, and maintainable updates across different IT environments.

Table 2: Core modules and technical realization of the prototype.

Architecture Layer	Main Realization	Role in Smart Failure Management
Frontend	HTML, CSS, Bootstrap, JavaScript	Responsive, user-friendly capture and interaction at workstation and desktop level
Application logic	Python Flask backend	Request handling, orchestration of analytics, role-dependent workflows
Data layer	PostgreSQL plus file storage	Persistent storage of structured and unstructured failure knowledge
AI/NLP services	Llama 3, Rasa, BERT/T5-based components	Conversational support, classification, guided questioning, semantic reasoning
Deployment layer	Docker containers	Portable, isolated, scalable deployment in heterogeneous industrial environments

On the presentation layer, the interface was developed with HTML, CSS, Bootstrap, and JavaScript. The frontend supports responsive usage on stationary and mobile devices and enables direct validation, asynchronous content loading, and low-threshold interaction with the assistance functions. On the server side, a Flask-based backend handles user requests, controls data processing, and connects classical data-analysis libraries such as Pandas, NumPy, and Scikit-learn with deep learning frameworks including PyTorch and TensorFlow.

The central persistence layer is a relational PostgreSQL database. According to the implementation concept, the database comprises seventeen core tables grouped into five major categories: failure data; product, process, and employee data; method-specific application data; interaction data from the analytical modules; and user and role management. This structure supports both operational traceability and higher-level analytical functions such as dashboards, similarity matching, and inference support.

The architecture also includes role-based access control. Administrators manage users and permissions, quality engineers curate and validate knowledge objects, and shop-floor personnel capture and process failure events. This differentiation is crucial for maintaining data quality and organizational accountability while keeping user interaction aligned with actual industrial responsibilities.

AI-ENABLED ASSISTANCE MECHANISMS

The novelty of the system lies in the integration of deterministic quality methods with probabilistic AI-based inference. Rather than treating AI as a black-box replacement for quality engineering, the system embeds AI where

it can increase speed, consistency, and accessibility without undermining explainability.

First, the architecture incorporates transformer-based NLP components for targeted sub-tasks. BERT-based classification is used to categorize root-cause statements, thereby supporting the systematic structuring of historical and newly recorded failure information. In parallel, a T5-based generation mechanism assists the 5-Why method by producing context-appropriate why-questions that help users progress from observed symptoms toward deeper causal structures.

Second, an LLM-supported conversational assistant enables contextual reasoning over both structured database entries and semi-structured historical documents. The assistant is built around a retrieval-augmented generation strategy. Historical reports are transformed into retrievable knowledge units and linked to semantic search mechanisms, so that the LLM receives domain-relevant context before generating answers. This reduces the risk of unsupported responses and improves alignment with company-specific failure knowledge.

Third, the system incorporates feedback loops. Suggested causes and corrective actions are not automatically accepted as truth. Instead, they are reviewed, refined, and validated by users. Validated outcomes are fed back into the knowledge base. This establishes a controlled learning cycle in which human expertise remains central, while AI increases retrieval quality, method accessibility, and interaction efficiency.

VALIDATION AND RESULTS

Validation was conducted in several stages in order to assess both conceptual suitability and practical effectiveness. Initial pre-validation involved experts who reviewed the conceptual design and the technical feasibility of the system. Subsequent validation comprised three technical test scenarios and a pilot deployment in industrial practice.

The first test scenario examined end-to-end system functionality in a real production environment at a steel company. The system's usability, process stability, and structured support of the problem-solving flow were positively assessed. The second scenario focused on the ingestion of historical failure documentation from an automotive supplier. Here, the automated derivation of causes and corrective actions showed high content agreement with the assessments of quality engineers. The third scenario compared system-supported case handling with a conventional approach in the context of Lean Six Sigma training. The group using the assistance system achieved better results in processing time, method confidence, and user satisfaction.

The strongest practical evidence came from the pilot deployment. The system was introduced at five workstations and embedded into real operating procedures. KPI-based evaluation showed a mean reduction of about 19 minutes in failure handling time, a lower recurrence tendency of comparable failures, and improved sustainability of corrective actions. Based on reduced personnel binding, annual savings of approximately EUR 60,940 were calculated for the pilot context. Qualitative feedback from employees and quality engineers further emphasized the value of the structured digital

guidance, especially for complex cases handled through the Berlin problem-solving cycle.

Notably, these results were obtained with a comparatively compact Llama 3 model with 8 billion parameters. This suggests that industrially viable performance does not necessarily depend on the largest available foundation models. Instead, domain grounding through RAG, careful workflow design, and method-based interaction can generate substantial practical value even under SME-compatible resource constraints.

DISCUSSION

The results indicate that smart failure management in SMEs benefits most when AI is embedded into a transparent methodological scaffold. In practice, the decisive challenge is not simply the availability of advanced models, but the conversion of fragmented failure information into structured, actionable knowledge. The proposed system addresses this by connecting method guidance, role-specific interaction, and AI-supported semantic processing in one architecture.

Several implications follow. First, the integration of PDCA-oriented quality methods with data-driven and language-based assistance can lower method-related barriers for less experienced users. Second, historical failure documents represent a high-value but underused asset; once structured and retrievable, they become a practical foundation for organizational learning. Third, explainability remains critical in regulated production contexts. The prototype therefore preserves the logic of established methods while using AI to accelerate retrieval, classification, and interaction rather than to bypass engineering judgement.

At the same time, limitations remain. The prototype was validated in selected industrial settings rather than across a large-scale multi-company rollout. Data quality and documentation maturity strongly influence the achievable performance of both semantic retrieval and machine learning components. Moreover, long-term effects on organizational learning, user acceptance, and governance of continuously expanding knowledge bases require further study.

CONCLUSION AND OUTLOOK

This paper presented the design and implementation of a modular, knowledge-based assistance system for SFM in manufacturing SMEs. The contribution lies in bringing together requirement-driven system design, method-guided quality engineering, historical knowledge reuse, and LLM-supported interaction within a single industrial assistance architecture.

The developed artifact demonstrates that SMEs can benefit from hybrid AI without abandoning the rigor of established quality management practice. Instead, deterministic methods such as 8D, FMEA, Ishikawa, and 5-Why can be digitally strengthened through semantic retrieval, guided questioning, and

explainable knowledge support. The validation results confirm the practical relevance of this approach and indicate measurable operational benefits.

Future work should investigate long-term learning effects, broader cross-company transferability, and stronger integration with enterprise systems. In addition, future versions may compare alternative embedding models, retrieval strategies, and larger or more specialized language models. The core insight nevertheless remains stable: the path toward intelligent failure management in SMEs lies not in replacing structured quality work, but in augmenting it with knowledge-centered and context-aware AI.

REFERENCES

Caglar, Turgut Refik, and Roland Jochem. 2023. "A System Approach for Creating Employee-Oriented Quality Control Loops in Production for Smart Failure Management System in SMEs." Intelligent Human Systems Integration (IHSI 2023): Integrating People and Intelligent Systems 69 (69). https://doi.org/10.54941/ahfe1002835.

Caglar, Turgut Refik, Elena Andrushchenko, Lennart Müller-Stein, and Roland Jochem. 2025a. "Systematische Entwicklung eines smarten Fehlermanagementsystems in der Produktion für KMU." In Rethinking Quality - Wandel des Qualitätsmanagements durch Digitalisierung und Künstliche Intelligenz, edited by Roland Jochem and Maurice Meyer. Springer Fachmedien. https://doi.org/10.1007/978-3-658-47213-9_12.

Caglar, Turgut Refik, Elena Andrushchenko, Lennart Müller-Stein, Christopher Wijayanto and Roland Jochem. 2025b. „ E-Learning as a Catalyst for Competence Development in Smart Failure Management: A SME-Focused Approach" 16th International Conference on Applied Human Factors and Ergonomics (AHFE 2025). https://doi.org/10.54941/ahfe1006272