Large Language Model-based Chatbot for Improving Human-Centricity in Maintenance Planning and Operations

Linus Kohl[1,2], Sarah Eschenbacher[1], Philipp Besinger[1] and Fazel Ansari[1,2]

[1]*Fraunhofer Austria Research GmbH, Center for Sustainable Production and Logistics, Vienna, 1040, Austria*
linus.kohl@fraunhofer.at
sarah.eschenbacher@fraunhofer.at
philipp.besinger@fraunhofer.at

[2]*TU Wien, Research Group of Production and Maintenance Management, Vienna 1040, Austria*
fazel.ansari@tuwien.ac.at

Abstract

The recent advances on utilizing Generative Artificial Intelligence (GenAI) and Knowledge Graphs (KG) enforce a significant paradigm shift in data-driven maintenance management. GenAI and semantic technologies enable comprehensive analysis and exploitation of textual data sets, such as tabular data in maintenance databases, maintenance and inspection reports, and especially machine documentation. Traditional approaches to maintenance planning and execution rely primarily on static, non-adaptive simulation models. These models have inherent limitations in accounting for dynamic environmental changes and effectively responding to unanticipated, ad hoc events.

This paper introduces a *maintenance chatbot* that enhances planning and operations, offering empathetic support to technicians and engineers, boosting efficiency, decision-making, and on-the-job satisfaction. It optimizes shift scheduling and task allocation by considering technicians' skills, physical stress, and psychological state, thus reducing cognitive stress. The approach ultimately improves human performance and reliability, embodying a human-centricity in the domain of maintenance and health management.

The practical impact of the *maintenance chatbot* is illustrated through its application in maintenance of railway cooling systems. The presented use case demonstrates the chatbot's potential as a transformative tool in maintenance management. Finally, the paper discusses the theoretical and practical considerations, in particular in the light of regulative frameworks such as EU AI ACT, highlighting the future pathways for complying with responsible AI requirements.

1. Introduction

The industrial landscape is currently facing a significant challenge due to the shortage of skilled labors, exacerbated by the increasing complexity of machinery and technological systems, as well as green transition, leading to limiting production by 28% in the European Union (EU) (European Commission 2023). This shortage poses a critical threat to the operational efficiency and sustainability of maintenance operations within various sectors. The complexity of modern machines requires a high level of expertise, yet industries often find themselves compelled to hire workers who may not fully meet these competency requirements (Shin et al. 2021). The European Union estimates the investment needed to reskill and upskill in manufacturing to 4.1 billion EUR up to 2030 (European Commission 2023). This gap between the required and available skill sets leads to inefficiencies, increasing human failure, thus reducing reliability and increasing downtime, and a greater potential for errors in maintenance operations.

Simultaneously, advancements in GenAI and semantic technologies have opened new avenues for capturing and leveraging the domain knowledge of experienced professionals (Abu-Rasheed et al. 2024), and at the same time assisting them on improving their problem-solving capabilities, e.g. through query-answers with chatbots (Kohl und Ansari 2023b). These technologies, particularly Large Language Models (LLMs), demonstrate an unparalleled capacity to analyze and interpret complex datasets, including technical documentation, maintenance logs, and operational reports (Birhane et al. 2023). Their ability to generate contextually relevant, accurate responses based on vast amounts of textual information marks a significant step forward in the development of cognitive assistants for maintenance tasks.

The intersection of skilled labor shortages (as a problem space) and GenAI technologies (as a solution space) underscores a critical need for tools that can bridge the gap between the complexity of modern machinery and the competencies of the available workforce. Cognitive assistance in maintenance, facilitated by AI-driven solutions, offers a promising approach to address this challenge (Kohl und Ansari 2023a). By providing real-time, tailored information and support, such tools can enhance decision-making, reduce cognitive load, and improve the efficiency of maintenance technicians who may not possess the full spectrum of required competencies and experiences. Furthermore, the integration of GenAI and semantic technologies in maintenance operations enables the preservation and dissemination of expert knowledge, mitigating the risk of knowledge loss due to workforce turnover or the retirement of seasoned professionals (Alavi et al. 2024). This capability is particularly valuable in light of the increasing complexity and specificity of modern industrial systems, where the loss of domain-specific knowledge can have significant operational impacts (Ansari 2019).

The need for cognitive assistance in maintenance is not only a response to the skilled labor shortage but also a strategic investment in the quality and reliability of maintenance operations. By enhancing the capabilities of maintenance technicians, engineers and planners, AI-driven tools can contribute to more resilient, efficient, and effective maintenance practices. The development and implementation of such tools, as exemplified by the LLM-based maintenance chatbot presented in this paper, represent a forward-looking approach to addressing the challenges of the contemporary industrial maintenance landscape (Romero und Stahre 2021). The following paper addresses the challenge of improving the workflow of maintenance operations and planning by leveraging LLM and semantic information.

The rest of the paper is structured as follows: In Section 2, the state-of-the-art is described, focusing on cognitive assistance system, Generative AI, especially Large Language Models. Thus, the research gap is identified. Section 3 introduces the system architecture and modular chatbot design, and Section 4 elaborates on its use case. Finally, Section 5 discusses the key findings and identifies the pathways for future research.

2. State of the art

This section explores the capabilities and applications of cognitive assistance systems within industrial manufacturing, emphasizing their role in augmenting human capabilities. It highlights how these systems utilize advanced technologies such as LLMs and KGs to optimize task execution. Additionally, it addresses the implications of the EU AI Act, which mandates transparency and safety in the deployment of such AI-driven systems, ensuring their responsible application in industrial environments.

2.1. Cognitive assistance system

Digital assistance systems (DAS) support workers in production, assembly and logistics to carry out their tasks efficiently in line with the situation and context (Ansari et al. 2020). These systems facilitate tasks ranging from scheduling and information retrieval to more complex operations, leveraging user inputs to deliver relevant outcomes and insights (Pokorni und Constantinescu 2021). Cognitive assistance systems (CAS), particularly within the manufacturing sector, extend this concept by focusing on augmenting human capabilities in intricate tasks rather than substituting human efforts (Kernan Freire et al. 2023), which can draw conclusions from its experience on the basis of significant portions of suitably presented knowledge so that it provides more appropriate, accurate or up-to-date information in its next use. These systems are engineered to support complex activities, including lifelong learning (Freire et al. 2023), machine operation, and task execution, through advanced methods of human-machine interaction (Listl et al. 2021). Employing a broad spectrum of techniques such as natural language processing (NLP) (Ansari et al. 2021), pose estimation for ergonomic risk identification (Kostolani et al. 2022), perception, and augmented reality (Zigart und Schlund 2020), CAS are designed to foster an intuitive and efficient interface for users.

CAS, utilizing NLP for natural language understanding, generation, and dialogue management, represent the most widespread interaction modality within CAS (Kang et al. 2020). These CAS are capable of engaging users in meaningful conversations, thereby facilitating labor-intensive tasks across multiple sectors, including customer service, healthcare, education, and manufacturing, through efficient and reliable communication (Eloundou et al. 2023).

In the industrial context, the application of CAS is an evolving research domain with significant potential benefits (Mark et al. 2021). These include providing centralized access to diverse information systems, decision making (Rožanec et al. 2022), delegating tasks (Burggräf et al. 2021), and enabling hands-free and gaze-free interactions (Romero und Stahre 2021), thereby enhancing operational efficiency and safety. Additionally, CAS in manufacturing can serve as valuable tools for on-the-job training (Wang et al. 2022) and real-time machine parameter adjustments, thereby contributing to the flexibility and adaptability of manufacturing processes (Zheng et al. 2022). Such applications highlight the transformative potential of cognitive assistants in augmenting human work, optimizing task execution, and facilitating continuous learning and adaptation in complex industrial environments.

2.2. Generative AI and Large Language Models

According to the OECD, GenAI “creates new content in response to prompts, offering transformative potential across multiple sectors such as education, entertainment, healthcare and scientific research”(OECD Artificial Intelligence Papers 2024). It, therefore, significantly broadens AI's application spectrum (Gozalo-Brizuela und Garrido-Merchan 2023). At the heart of GenAI's advancements are LLMs like Generative Pre-trained Transformers (GPT), which have dramatically enhanced AI's language processing and generating capabilities, offering applications from automating documentation to improving decision-making in industries.

Retrieval-Augmented Generation (RAG) (Jing et al. 2024) extends LLMs by integrating them with information retrieval systems, enabling real-time access to extensive databases for more precise, context-specific causal outputs (Zhou et al. 2024). This is particularly valuable in manufacturing and maintenance, where accessing up-to-date technical and diagnostic information is crucial (Kernan Freire et al. 2023).

AI agents represent a further advancement, capable of autonomous decision-making based on environmental learning and adaptation (Zhao et al. 2023). In the context of manufacturing and maintenance, these agents can autonomously monitor system health (Han und Tao 2024), predict (Saboo und Shekhawat 2024) and automate maintenance tasks (Sun et al. 2024), thereby reducing downtime and maintenance costs. It can therefore be said that current approaches can achieve relevant results through their purely probabilistic transformer architecture by using attention with classical RAG, but cannot use factual, linked knowledge.

2.3. Knowledge Graph

Knowledge Graphs (KGs) structure knowledge in graphs, connecting entities and their relationships, thereby facilitating semantic searches and data integration (Fensel et al. 2020). In GenAI applications, KGs enhance RAG (Zhu et al. 2024) by providing structured, semantically linked domain information to improve response accuracy and contextual relevance (Agarwal et al. 2020), particularly valuable in domain-specific applications like manufacturing (Yu 2022). KG therefore enhance capabilities of chatbots by providing them with structured context information on specific user requests (Li et al. 2021). By leveraging the rich semantic relationships within KGs, chatbots are able to understand and process user queries more effectively, navigating through complex information networks to retrieve or infer accurate answers (Yu 2021).

Within manufacturing, KGs encapsulate domain knowledge and causal relationships between failure modes and solutions, informed by Failure Modes and Effects Analysis (FMEA) (Razouk et al. 2023). This structured knowledge aids RAG systems in querying precise information for predictive maintenance and decision support, thereby streamlining maintenance protocols and diagnosing machinery issues through an understanding of causal links.

The synergy between KGs and RAG significantly enhances manufacturing operations' efficiency by enabling access to detailed domain knowledge, reducing downtime, and guiding accurate maintenance decisions, thus enhancing operational reliability and performance (Ansari et al. 2023).

2.4. EU AI Act

In response to the rapid developments in the field of AI in recent years, the European Union has implemented a regulatory framework for development, market introduction and deployment of AI-driven products, services, and systems. The framework is designed to guarantee transparency, accountability, and safety for both current and forthcoming AI technologies within the EU. Especially in the area of manufacturing a responsible application of AI is essential to mitigate risks and deliver business benefits (Besinger et al. 2024).

Since current pre-trained LLMs like the GPT-models (Brown et al. 2020) or Metas Llama-Series (Touvron et al. 2023) are trained outside of the European Union, the EU AI Act addresses this issue by extending its scope to include providers operating within the EU as well as those in third countries, particularly when the output of their AI systems is utilized within the Union. The EU AI Act defines different categories from no risk to high-risk. The use of AI in human interaction, emotion recognition, and content generation is categorizes as low risk (second category). Article 52 (European Commission 2024) addresses the regulatory requirements for providers and users (excluding end-users) of AI systems categorized as low risk. There are three critical areas pertinent to the case presented in this paper: Transparency in AI interactions, the Marking of synthetic content, and the Disclosure requirements for emotion recognition and biometric categorization.

Firstly, concerning Transparency in AI interactions, the legislation mandates that AI systems engaging in human interaction must inform users of their non-human nature, except in contexts where such interaction is inherently apparent. Secondly, the requirement for marking synthetic content, such as audio, images, videos, or text, created or significantly altered by AI there must be machine-readable marks signifying its artificially generated or manipulated status, except for minor edits. Lastly, concerning emotion recognition and biometric categorization, users must be informed about these processes, with data handling needing to comply with EU regulations (European Commission 2024).

2.5. Research Gap

In industrial maintenance, the accessibility and quality of critical data is a crucial issue. Despite the increasing availability of information from maintenance reports, personnel documents, and enterprise resource planning (ERP) systems, the effective use of this data remains largely untapped. Therefore, to the best of the authors' knowledge, current research has not sufficiently explored the use of LLM in chatbots in industrial applications, especially the use of linked data and documents in an agent network. This paper presents a novel way to combine LLM with RAG and KG in an intent-driven agent framework, providing a flexible, generalizable and scalable approach for industrial maintenance.

3. METHODOLOGY

In the following the design of the system architecture, which enables an LLM-based maintenance chatbot is described. Further, we propose a modular agent layout for the chatbot. The architecture is based on by RAMI 4.0 (DIN 91345) and inspired by (Margaria und Schieweck 2019).

3.1. System Architecture

The system architecture for the application of an LLM-based maintenance chatbot, see Figure 1. is structured into three distinct tiers, namely data tier, analytics tier and presentation tier, each with specific components, capabilities and information flows designed to interact seamlessly within the broader ecosystem of the industrial application.

Data Tier: The *Event Broker* facilitates communication between the analytics components and the data sources. It manages the flow of real-time data to the Data Analytics (Stream) and routes information to and from the Prescriptive Analytics. The *Database* stores historical data, such as CAD-models, maintenance reports or technical data, which is subsequently used for trend analysis and informing predictive models. It also serves as a repository for collected data over time and connects them through semantic similarities, which leverages the suitability of natural language interaction. Vectorized data schemas in the Data Tier allow for efficient data retrieval. *Edge Devices* are directly connected to the database and serve as intermediaries between the physical sensors and the system's core data infrastructure. They perform preliminary data processing, filtering, and aggregation tasks.

Sensors, either attached to machines or environmental sensors, collect data about the operational status, health, and performance of the machinery and environmental status. This data is crucial for monitoring and maintenance purposes. It incorporates different database structures. For processing natural language, the core components are a vector database and a KG, which serve as the foundation for an efficient RAG pipeline. While vector databases enable efficient data retrieval through vectorized representation of domain specific data (Jing et al. 2024), KGs provide structured representation of the data (Pan et al. 2024). A combination of these components is leveraged to reduce hallucination and utilize information which is not inherent to the LLM. The KG, see Figure 1, enables the connection of ERP data with task and competence relevant information. This data model allows a holistic view on the maintenance process as well as the possibility for downstream agents for interconnected reasoning. *Machines* are the physical hardware being monitored and maintained. Connected to sensors, they are the source of the operational data fed into the system for analysis. *Assistance Systems*, such as smart tools and tablets, are connected to sensors. They serve as an interface for workers

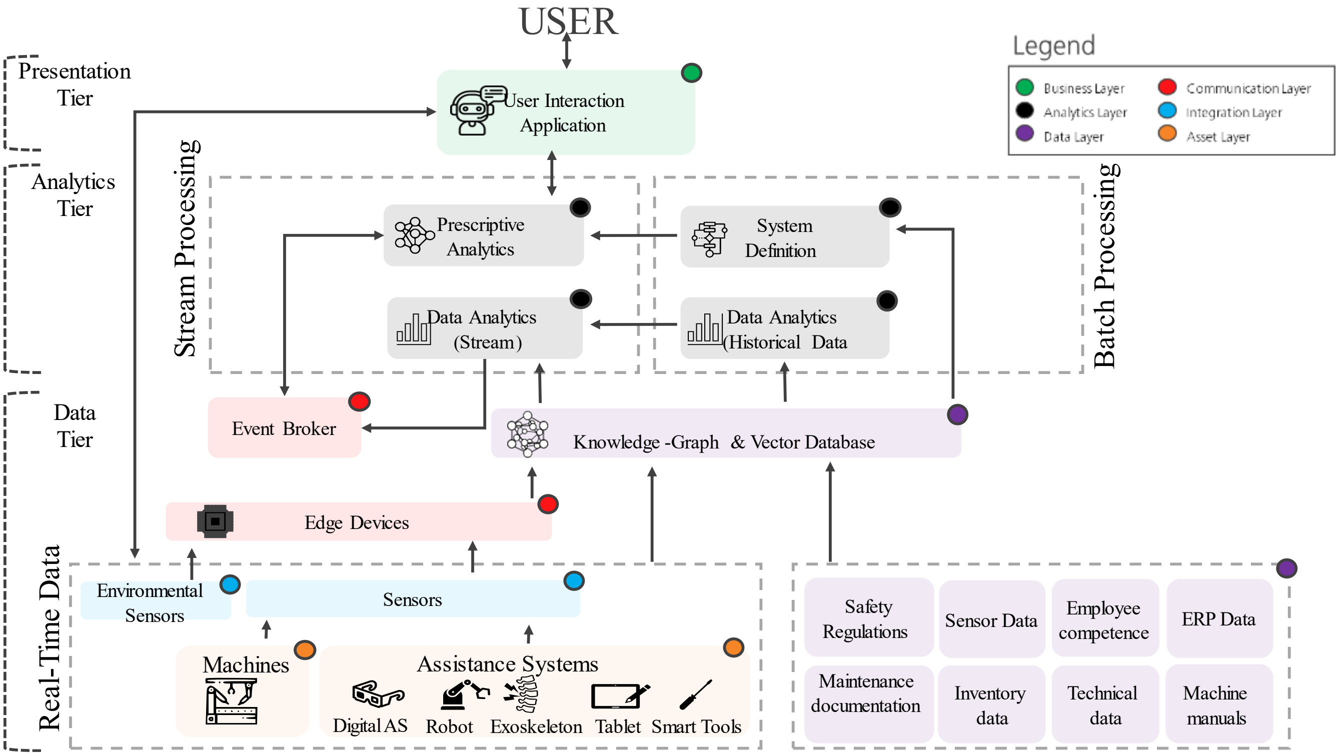


Figure 1: System Architecture Layout for empathetic assistance systems

on the ground, providing them with real-time guidance derived from the system's analysis.

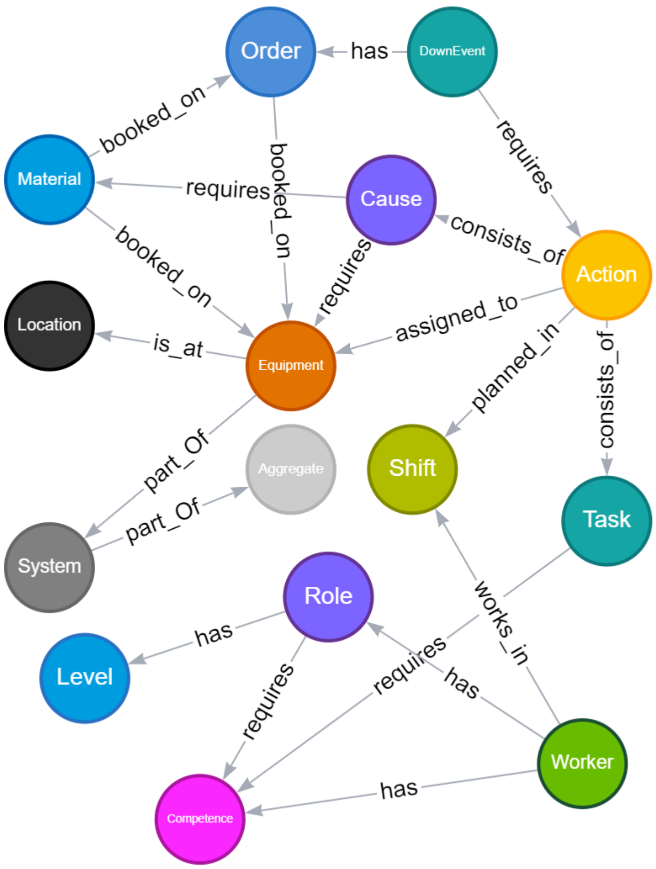


Figure 2: Data model of the KG based extended from (Kohl und Ansari 2023a)

Analytics Tier: The *Prescriptive Analytics* component is directly linked to the User Interaction Application. It processes the user's input, utilizing the Rasa conversational AI framework (Introduction to Rasa Open Source & Rasa Pro 2024), to generate actionable advice or maintenance recommendations. It uses advanced algorithms such as anomaly detection to suggest specific actions based on the analyzed data. The *System Definition* , powered by the Llama-2-70b-model (Touvron et al. 2023), functions as a reasoning framework that defines and orchestrates data analytic processes. It incorporates a multi-agent layer structure to process user input and determine the most appropriate action to take (Jiang et al. 2023). Therefore, necessary parameters and fitting data sources are determined to resemble the scope for aspired data analytics. The *Data Analytics (Historical Data)* component uses batch processing to analyze historical data to identify trends, patterns and potential issues based on past events. In contrast to the historical data analytics, the *Data Analytics (Stream)* component processes simulated real-time sensor data to offer immediate insights and detect current or impending issues, which is essential for real-time decision-making and alerts. The Data Analytics component utilizes multiple regression to forecast outcomes and incorporates K-means clustering to discover trends in historical data, as well as an Isolation Forest algorithm for anomaly detection. The foundational understanding of the Data Analytics (Historical Data) additionally augments the predictive real-time models to ensure a maximum of information for analysis.

Presentation Tier: The *User Interaction Application* component serves as the interface between the end-user and the chatbot system. It is where users interact with the chatbot, inputting queries and receiving responses. In this context, a simplistic User Interface featuring a Chatbot window was implemented, as illustrated in Figure 4.

Each tier in this architecture is intricately connected, allowing data to flow from the machines up through the system to enable real-time and predictive maintenance decision-making. The architecture is designed to maximize efficiency, reduce mean time to repair, and provide actionable insights through a user-friendly interface.

3.2. Modular Chatbot Layout

This chatbot layout is aligned with existing frameworks for developing multi-agent dialogue systems (Engelmann et al. 2023; Xi et al. 2023). It is depicted in Figure 3 and features a central User Agent linked to three specialized agents (Scheduling, Competency, Analyzer), all interfacing with an

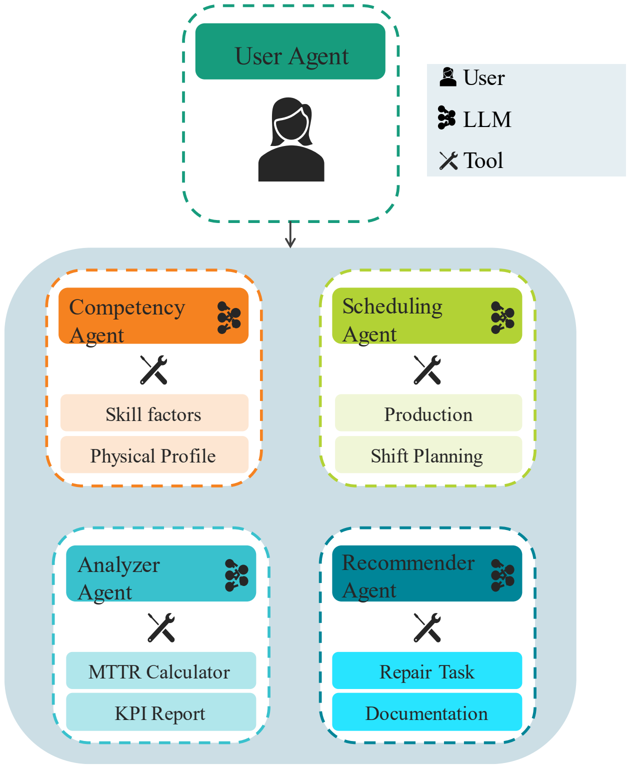


Figure 3: Interconnected Agent Layout for a modular maintenance chatbot architecture

LLM acting as a classification engine to determine which agent is triggered for a certain query. This User Agent represents the System Definition within the System Architecture, see Figure 1 and therefore determines which specialized agent is triggered subsequently. These specialized agents comprise of several tools and determine the correct tool usage for task-specific challenges. Moreover, the agents can interface with each other if the task requires agent collaboration. The proposed modular design allows for seamless integration of further agents and tools within agents to encounter novel challenges over time.

- **User Agent:** Channels user inputs to the appropriate specialized agents and consolidates their outputs for user communication. It is the link between Presentation and Analytics Tier.
- **Scheduling Agent:** Selects the production planning and shift planning tools based on user agent task instructions and operational needs. The optimization algorithm leverages provided data sources within the Data Tier and interacts with the Competency and Analyzer Agent to ensure fairness and efficiency while allocating shifts and schedule production.
- **Competency Agent:** Decides whether to analyze skill factors or physical profiles, aligning workforce tasks with individual skills and physical capabilities for optimal job assignment. Through its empathetic capability it continuously checks for physical and ethical alignment of worker tasks.
- **Analyzer Agent**: Chooses between MTTR calculation and KPI report analysis tools to assess maintenance effectiveness and identify areas for operational improvement. It provides recommendations such as prioritization or suggestions for automations.
- **Recommender Agent:** Has access to both historical and real-time data. When an anomaly is detected, it becomes operational. It offers similar historical failures, spare parts, and can store documentation in the KG.

Contrary to the User Agent the specialized agents interact with the Data Tier and leverage aforementioned RAG pipelines with KGs and vector databases to process dynamic and real-time information (Huang et al. 2024). This layout serves as an illustration of how agents can be utilized to allow dynamic maintenance strategies. The system architecture, see Figure 1, provides a high-level reference structure for the integration of new agents, such as a failure mode and effects analysis (FMEA) agent using the cause entity from the KG.

The architecture of this modular system integrates prompts as follows: The overarching system prompt guides the Chatbot, setting its function within a maintenance environment. This structure includes more specific prompts at subordinate levels. The User-Agent prompt functions analogously to a supervisory agent, tasked with identifying the most appropriate agent response to a user query. Each specialized agent operates under its own prompt; for example, the Analyzer Agent is responsible for generating reports based on historical or real-time data. This necessitates determining whether to initiate tools such as MTTR or KPI reports. Subsequently, this agent classifies the tool required for the task, parsing input parameters – such as the specific machine and time span – from the LLM. These parameters, where descriptions are also provided, are then employed within Python functions, with the resulting outputs fed back to the LLM, which then crafts responses based on these function outputs.

3.3. Regulative Considerations

In the context of implementing a maintenance chatbot, aligning with the EU AI Act's transparency obligations is essential for fostering trust and ensuring responsible use. The EU AI Act mandates that users are explicitly informed when they are interacting with AI systems, like chatbots. This requirement is critical in maintenance environments, where decisions can impact operational safety and efficiency. By disclosing the chatbot's AI nature, users are empowered to make informed choices about their engagement, understanding that they are consulting a machine for assistance. This transparency not only builds trust in the technology's capabilities and limitations but also reinforces the importance of human oversight in decision-making processes. Ensuring users are aware they are interacting with

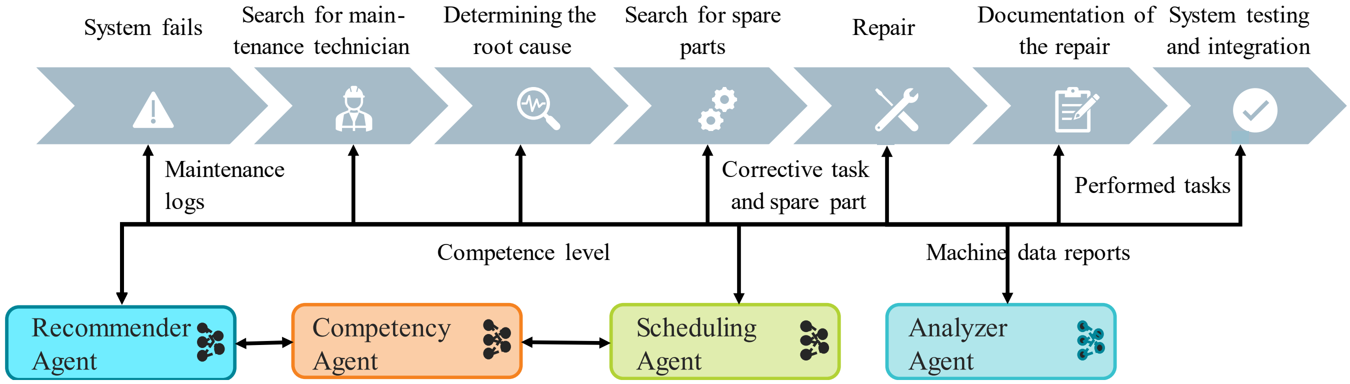
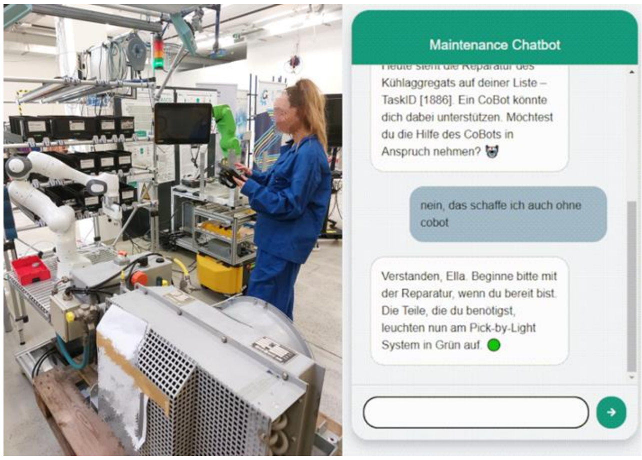


Figure 4: Maintenance process and its triggered agents as well as information flow

an AI helps maintain a balance between leveraging technological advancements and preserving human judgement and accountability in maintenance operations.

4. USE-CASE

The use case discusses a maintenance workflow in the railway industry, utilizing a chatbot for maintaining a cooling system., see Figure 4. The used cooling system provides sensor data about machine states as well as extensive manufacturer information. As it is one of the most frequently installed systems in Vienna's tramways and subways, extensive data on maintenance incidents in form of logs and spare parts is available. The data tier and the used maintenance data set consists of text-based, tabular industrial maintenance logs (Ansari 2020). The dataset was transformed in order to fit the structure of the Sequential QA (SQA) format by Microsoft (Iyyer et al. 2016) in order the be ideally processed by LLMs. In the use case scenario, the KG is constructed from maintenance logs exported from an ERP system, detailing machine failures and corrective actions, also integrates information on the equipment and spare parts used for repairs. The association of actions with required competence and the frequency of these actions by maintenance technicians serve to depict their competence levels (Ansari et al. 2023). Further, a vector database houses segments, specifically text excerpts, from work instructions and machinery documentation. For real-time data, the system monitors the current production schedule along with a simulated data stream of sensor readings from the machines. Additional stress levels of maintenance technicians are recorded for evaluation purpose.



Heute steht die Reparatur des Kühlaggregats auf deiner Liste – TaskID [1886]. Ein CoBot könnte dich dabei unterstützen. Möchtest du die Hilfe des CoBots in Anspruch nehmen?

nein, das schaffe ich auch ohne cobot

Verstanden, Ella. Beginne bitte mit der Reparatur, wenn du bereit bist. Die Teile, die du benötigst, leuchten nun am Pick-by-Light System in Grün auf. ●

Figure 5: Maintenance of a railway cooling system using a chatbot

4.1. Application of the Chatbot

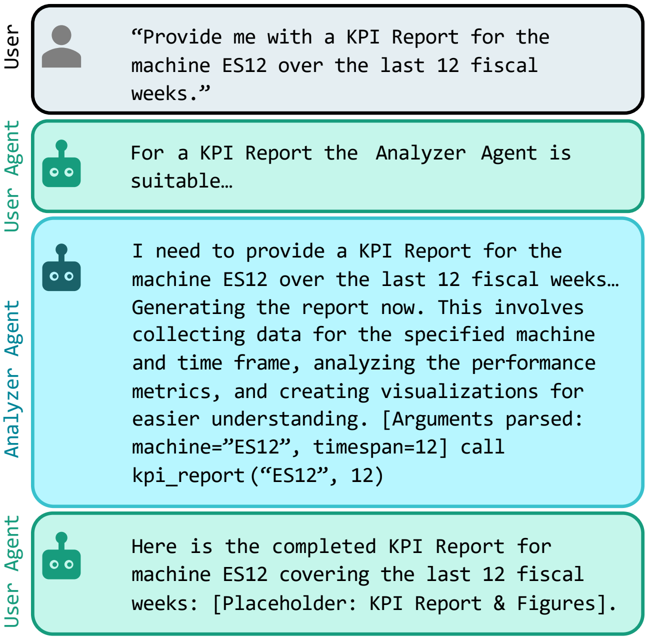
The chatbot's supportive capability is discussed based on the standard end-to-end maintenance process see Figure 4. It consists of equipment failure, search for maintainer, identification of fault cause, search for spare parts, repair action, documentation of the maintenance process, reintegration of the machine. The following shows points of human interaction as well as autonomous chatbot within this process.

1. **Equipment failure**: The recommender agent is activated by an error notification, triggered by an anomaly in the real time data flow of the machine. Based on the error notification similar historical failures and corresponding actions are determined by semantic search of the task recommendation agent (Ansari et al. 2021).
2. **Search for maintainer**: The scheduling agent, competency agent and task recommendation agent exchange information about the production schedule, available maintenance personnel, their corresponding competencies, and the necessary tasks for failure resolution. According to that an allocation of the most fitting maintainer for the task is deducted.
3. **Identification of fault cause**: This stage marks the initial interaction between the maintainer and the chatbot. Utilizing the chatbot's knowledge, sourced from documents within the vector database, it can pose inquiries related to specific domains or machinery. Throughout this process, the human evaluates the tasks recommended by the chatbot for accuracy and corroborates them based on personal experience and the information furnished by the chatbot.
4. **Search for spare parts**: Once the tasks required for resolving the failure are identified, the task recommendation agent traverses through historical data in the KG to propose necessary spare parts.
5. **Repair action**: During the physical repair, the chatbot acts as an accessible source of pertinent information, offering guidance through machine documents or other necessary data from the vector database. Additionally, it can process requests for more detailed machine information, which are then thoroughly examined by the analyzer agent, e.g. asking for the mean time of repair.
6. **Documentation of maintenance process**: Building on prior interactions, the chatbot can autonomously create new connections within the KG and carry out the documentation process upon request from the maintenance personnel.
7. **Reintegration of the machine**: Finally, the chatbot guides through standard tasks to reintegrate the machine levering information from diverse work instructions.

The proposed integration of a chatbot within the standard end-to-end maintenance process, see Figure 4, represents a significant advancement in operational efficiency and precision. By embedding intelligent, autonomous capabilities at critical junctures of the maintenance workflow, from initial equipment failure detection to the reintegration of repaired machinery, this model showcases a transformative shift towards more resilient and adaptive maintenance operations. The synergy between human expertise and artificial intelligence not only enhances the decision-making process but also optimizes resource allocation, reduces downtime, and enables empathic human-machine collaboration (Sorin et al. 2023).

4.2. Example: Analyzer Agent

To illustrate a potential maintenance workflow, we present an example of a chat where a user activates the Analyzer Agent. The dialogue demonstrates the triggered process chains, including the reasoning (highlighted in italics) and the tools employed. Due to space constraints, a detailed description of the various prompts is largely omitted. The chatbot is referred to as "AI." Users can also follow the thought process through a collapsible dropdown segment in the interface, see Figure 6.



"Provide me with a KPI Report for the machine ES12 over the last 12 fiscal weeks."

For a KPI Report the Analyzer Agent is suitable…

I need to provide a KPI Report for the machine ES12 over the last 12 fiscal weeks… Generating the report now. This involves collecting data for the specified machine and time frame, analyzing the performance metrics, and creating visualizations for easier understanding. [Arguments parsed: machine="ES12", timespan=12] call kpi_report("ES12", 12)

Here is the completed KPI Report for machine ES12 covering the last 12 fiscal weeks: [Placeholder: KPI Report & Figures].

Figure 6: Example prompt of the Analyzer Agent

This example highlights how the chatbot, in its current state, bases its decisions, extending to the collaboration of multiple agents to optimize outcomes, such as in production planning.

4.3. Evaluation of the maintenance-chatbot

The evaluation is based on two types of maintenance tasks performed in TU Wien's pilot factory: a simple task for changing and cleaning a filter and a more complex task for changing the rotor, where the root cause is not clear. The depicted tasks require different competence levels in different areas. In the test scenarios, the maintenance chatbot demonstrated promising results for guiding the maintenance technicians through the root cause identification and for offering more detailed answers when needed, thereby reducing MTTR by 25% in comparison to the control group. In the case of the more challenging problem setting, the possibility of creating KPIs in natural language for deeper analysis reduced the MTTR by approximately 30% leading to an even higher impact. Furthermore, the dialogues are tailored to the individual competence levels, which permit queries and elucidations of (partial) steps, diminish the stress level and cognitive load, and facilitate a more empathetic conversational style. In addition, a ground truth dataset was constructed. Based on the logs, the appropriate agent or tool was triggered, and that the response was satisfactory in 83% of all cases. Specific tools, such as LangSmith (Ito et al. 2020), are currently under evaluation concerning integration for even better agent handling. Given the implementation of the LLaMA2 model within this chatbot, the Do-Not-Answer dataset (Wang et al. 2023) establishes a framework for safeguarding LLMs against potential risks. The efficacy of this dataset in mitigating harms will be further assessed in forthcoming studies through an adapted version tailored to evaluate the specific vulnerabilities and challenges posed by this chatbot.

5. Conclusion and Outlook

In summary, this investigation highlights a maintenance chatbot's significant efficiency over traditional systems in minimizing Mean Time to Repair (MTTR), thereby boosting operational efficiency and equipment effectiveness in manufacturing. Traditional NLP based systems show an improvement in MTTR of at least 20% in production environments, which is confirmed by preliminary investigations by (Ansari et al. 2023). Additionally independent studies on LLM based assistance systems (Noy und Zhang 2023) show even higher potentials in operational efficiency, well in line with the first tests of the detailed maintenance chatbot in the pilot factory use cases. Leveraging advanced NLP and machine learning, the chatbot surpasses conventional systems by integrating ERP data and identifying relationships for enhanced maintenance insights, significantly reducing cognitive load and stress.

Looking ahead, the scalability and generalizability of the maintenance chatbot are poised for improvement with the multi-agent systems, and causal AI. AutoGen frameworks (Wu et al. 2023) are anticipated to refine the chatbot's content generation and adaptation capabilities, enabling reciprocal learning. Multi-agent systems promise to distribute problem-solving tasks effectively, improving maintenance operations' efficiency. Meanwhile, causal AI could provide a deeper understanding of the complex causal relationships within maintenance data systems, offering more accurate step-by-step solutions.

Future directions indicate that maintenance chatbots could overcome current limitations and adapt across various manufacturing settings. This flexibility is key to meeting the sector's varied needs, marking a significant advancement in CAS for maintenance. Driven by improvements in data integration, natural language processing, and causality understanding, this represents a crucial step in manufacturing's digital transformation.

ACKNOWLEDGEMENT

The authors would like to acknowledge the financial support of this research work by the Fraunhofer-Gesellschaft as part of the Fraunhofer flagship project EMOTION.

REFERENCES

Abu-Rasheed, Hasan; Abdulsalam, Mohamad Hussam; Weber, Christian; Fathi, Madjid (2024): Supporting Student Decisions on Learning Recommendations: An LLM-Based Chatbot with Knowledge Graph Contextualization for Conversational Explainability and Mentoring. Online verfügbar unter http://arxiv.org/pdf/2401.08517v3.

Agarwal, Oshin; Ge, Heming; Shakeri, Siamak; Al-Rfou, Rami (2020): Knowledge Graph Based Synthetic Corpus Generation for Knowledge-Enhanced Language Model Pre-training. Online verfügbar unter http://arxiv.org/pdf/2010.12688v2.

Alavi, Maryam; Leidner, Dorothy E.; Mousavi, Reza (2024): Knowledge Management Perspective of Generative Artificial Intelligence. In: *JAIS* 25 (1), S. 1–12. DOI: 10.17705/1jais.00859.

Ansari, Fazel (2019): Knowledge Management 4.0: Theoretical and Practical Considerations in Cyber Physical Production Systems. In: *IFAC-PapersOnLine* 52 (13), S. 1597–1602. DOI: 10.1016/j.ifacol.2019.11.428.

Ansari, Fazel (2020): Cost-based text understanding to improve maintenance knowledge intelligence in manufacturing enterprises. In: *Computers & Industrial Engineering* 141, S. 106319. DOI: 10.1016/j.cie.2020.106319.

Ansari, Fazel; Glawar, Robert; Nemeth, Tanja (2019): PriMa: a prescriptive maintenance model for cyber-physical production systems. In: *International Journal of Computer Integrated Manufacturing* 32 (4-5), S. 482–503. DOI: 10.1080/0951192X.2019.1571236.

Ansari, Fazel; Hold, Philipp; Khobreh, Marjan (2020): A knowledge-based approach for representing jobholder profile toward optimal human–machine collaboration in cyber physical production systems. In: *CIRP Journal of Manufacturing Science and Technology* 28, S. 87–106. DOI: 10.1016/j.cirpj.2019.11.005.

Ansari, Fazel; Kohl, Linus; Giner, Jakob; Meier, Horst (2021): Text mining for AI enhanced failure detection and availability optimization in production systems. In: *CIRP Annals* 70 (1), S. 373–376. DOI: 10.1016/j.cirp.2021.04.045.

Ansari, Fazel; Kohl, Linus; Sihn, Wilfried (2023): A competence-based planning methodology for optimizing human resource allocation in industrial maintenance. In: *CIRP Annals* 72 (1), S. 389–392. DOI: 10.1016/j.cirp.2023.04.050.

Besinger, Philipp; Vejnoska, Daniel; Ansari, Fazel (2024): Responsible AI (RAI) in Manufacturing: A Qualitative Framework. In: *Procedia Computer Science* 232, S. 813–822. DOI: 10.1016/j.procs.2024.01.081.

Birhane, Abeba; Kasirzadeh, Atoosa; Leslie, David; Wachter, Sandra (2023): Science in the age of large language models. In: *Nat Rev Phys* 5 (5), S. 277–280. DOI: 10.1038/s42254-023-00581-4.

Brown, Tom B.; Mann, Benjamin; Ryder, Nick; Subbiah, Melanie; Kaplan, Jared; Dhariwal, Prafulla et al. (2020): Language Models are Few-Shot Learners. Online verfügbar unter http://arxiv.org/pdf/2005.14165v4.

Burggräf, Peter; Dannapfel, Matthias; Adlon, Tobias; Föhlisch, Nils (2021): Adaptive assembly systems for enabling agile assembly – Empirical analysis focusing on cognitive worker assistance. In: *Procedia CIRP* 97, S. 319–324. DOI: 10.1016/j.procir.2020.05.244.

Eloundou, Tyna; Manning, Sam; Mishkin, Pamela; Rock, Daniel (2023): GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. Online verfügbar unter http://arxiv.org/pdf/2303.10130v5.

Engelmann, Débora C.; Panisson, Alison R.; Vieira, Renata; Hübner, Jomi Fred; Mascardi, Viviana; Bordini, Rafael H. (2023): MAIDS — A Framework for the Development of Multi-Agent Intentional Dialogue Systems. In: *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '23)*, S. 1209–1217.

European Commission (2023): Employment and social developments in Europe 2023. Luxembourg: Publications Office of the European Union (Employment and social developments in Europe 2023).

European Commission (2024): EU AI Act. Article 52, Transparency Obligations for Providers and Users of Certain AI Systems and GPAI Models. Online verfügbar unter https://www.euaiact.com/article/52, zuletzt geprüft am 27.03.2024.

Fensel, Dieter; Şimşek, Umutcan; Angele, Kevin; Huaman, Elwin; Kärle, Elias; Panasiuk, Oleksandra et al. (2020): Introduction: What Is a Knowledge Graph? In: Dieter Fensel, Umutcan Şimşek, Kevin Angele, Elwin Huaman, Elias Kärle, Oleksandra Panasiuk et al. (Hg.): Knowledge Graphs. Cham: Springer International Publishing, S. 1–10.

Freire, Samuel Kernan; Panicker, Sarath Surendranadha; Ruiz-Arenas, Santiago; Rusák, Zoltán; Niforatos, Evangelos (2023): A Cognitive Assistant for Operators: AI-Powered Knowledge Sharing on Complex Systems. In: *IEEE Pervasive Comput.* 22 (1), S. 50–58. DOI: 10.1109/MPRV.2022.3218600.

Gozalo-Brizuela, Roberto; Garrido-Merchan, Eduardo C. (2023): ChatGPT is not all you need. A State of the Art Review of large Generative AI models. Online verfügbar unter http://arxiv.org/pdf/2301.04655v1.

Han, Yu; Tao, Jingwen (2024): Revolutionizing Pharma: Unveiling the AI and LLM Trends in the Pharmaceutical Industry. Online verfügbar unter http://arxiv.org/pdf/2401.10273v2.

Huang, Xu; Liu, Weiwen; Chen, Xiaolong; Wang, Xingmei; Wang, Hao; Lian, Defu et al. (2024): Understanding the planning of LLM agents: A survey. Online verfügbar unter http://arxiv.org/pdf/2402.02716v1.

Introduction to Rasa Open Source & Rasa Pro (2024). Online verfügbar unter https://rasa.com/docs/rasa/, zuletzt aktualisiert am 22.03.2024.

Ito, Takumi; Kuribayashi, Tatsuki; Hidaka, Masatoshi; Suzuki, Jun; Inui, Kentaro (2020): Langsmith: An Interactive Academic Text Revision System. Online verfügbar unter http://arxiv.org/pdf/2010.04332v1.

Iyyer, Mohit; Yih, Wen-tau; Chang, Ming-Wei (2016): Answering Complicated Question Intents Expressed in Decomposed Question Sequences. Online verfügbar unter http://arxiv.org/pdf/1611.01242v1.

Jiang, Zhiqiu; Rashik, Mashrur; Panchal, Kunjal; Jasim, Mahmood; Sarvghad, Ali; Riahi, Pari et al. (2023): CommunityBots: Creating and Evaluating A Multi-Agent Chatbot Platform for Public Input Elicitation. In: *Proc. ACM Hum.-Comput. Interact.* 7 (CSCW1), S. 1–32. DOI: 10.1145/3579469.

Jing, Zhi; Su, Yongye; Han, Yikun; Yuan, Bo; Xu, Haiyun; Liu, Chunjiang et al. (2024): When Large Language Models Meet Vector Databases: A Survey. Online verfügbar unter http://arxiv.org/pdf/2402.01763v2.

Kang, Yue; Cai, Zhao; Tan, Chee-Wee; Huang, Qian; Liu, Hefu (2020): Natural language processing (NLP) in management research: A literature review. In: *Journal of Management Analytics* 7 (2), S. 139–172. DOI: 10.1080/23270012.2020.1756939.

Kernan Freire, Samuel; Foosherian, Mina; Wang, Chaofan; Niforatos, Evangelos (2023): Harnessing Large Language Models for Cognitive Assistants in Factories. In: Minha Lee, Cosmin Munteanu, Martin Porcheron, Johanne Trippas und Sarah Theres Völkel (Hg.): Proceedings of the 5th International Conference on Conversational User Interfaces. CUI '23: ACM conference on Conversational User Interfaces. Eindhoven Netherlands, 19 07 2023 21 07 2023. New York, NY, USA: ACM, S. 1–6.

Kohl, Linus; Ansari, Fazel (2023a): A Knowledge Graph-based Learning Assistance Systems for Industrial Maintenance, in press.

Kohl, Linus; Ansari, Fazel (2023b): Chatbots in der Instandhaltungsplanung: Industrielle Anwendungsfälle und zukünftige Perspektiven: ÖVIA Kongress.

Kostolani, David; Wollendorfer, Michael; Schlund, Sebastian (2022): ErgoMaps: Towards Interpretable and Accessible Automated Ergonomic Analysis. In: 2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS). 2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS). Orlando, FL, USA, 17.11.2022 - 19.11.2022: IEEE, S. 1–7.

Li, Yunqing; Raman, Shivakumar; Cohen, Paul; Starly, Binil (2021): Design of Knowledge Graph in Manufacturing Services Discovery. In: Volume 2: Manufacturing Processes; Manufacturing Systems; Nano/Micro/Meso Manufacturing; Quality and Reliability. ASME 2021 16th International Manufacturing Science and Engineering Conference. Virtual, Online, 21.06.2021 - 25.06.2021: American Society of Mechanical Engineers.

Listl, Franz Georg; Fischer, Jan; Weyrich, Michael (2021): Towards a Simulation-based Conversational Assistant for the Operation and Engineering of Production Plants. In: 2021 26th IEEE International Conference on Emerging

Technologies and Factory Automation (ETFA). 2021 IEEE 26th International Conference on Emerging Technologies and Factory Automation (ETFA). Vasteras, Sweden, 07.09.2021 - 10.09.2021: IEEE, S. 1–4.

Margaria, Tiziana; Schieweck, Alexander (2019): The Digital Thread in Industry 4.0. In: Wolfgang Ahrendt und Silvia Lizeth Tapia Tarifa (Hg.): Integrated Formal Methods, Bd. 11918. Cham: Springer International Publishing (Lecture Notes in Computer Science), S. 3–24.

Mark, Benedikt G.; Rauch, Erwin; Matt, Dominik T. (2021): Worker assistance systems in manufacturing: A review of the state of the art and future directions. In: *Journal of Manufacturing Systems* 59, S. 228–250. DOI: 10.1016/j.jmsy.2021.02.017.

Noy, Shakked; Zhang, Whitney (2023): Experimental evidence on the productivity effects of generative artificial intelligence. In: *Science (New York, N.Y.)* 381 (6654), S. 187–192. DOI: 10.1126/science.adh2586.

OECD Artificial Intelligence Papers (2024).

Pan, Shirui; Luo, Linhao; Wang, Yufei; Chen, Chen; Wang, Jiapu; Wu, Xindong (2024): Unifying Large Language Models and Knowledge Graphs: A Roadmap. In: *IEEE Trans. Knowl. Data Eng.*, S. 1–20. DOI: 10.1109/TKDE.2024.3352100.

Pokorni, Bastian; Constantinescu, Carmen (2021): Design and Configuration of Digital Assistance Systems in Manual Assembly of Variant-rich Products based on Customer Journey Mapping. In: *Procedia CIRP* 104, S. 1777–1782. DOI: 10.1016/j.procir.2021.11.299.

Razouk, Houssam; Liu, Xing Lan; Kern, Roman (2023): Improving FMEA Comprehensibility via Common-Sense Knowledge Graph Completion Techniques. In: *IEEE Access* 11, S. 127974–127986. DOI: 10.1109/ACCESS.2023.3331585.

DIN 91345, 2016: Referenzarchitekturmodell Industrie 4.0 (RAMI4.0).

Romero, David; Stahre, Johan (2021): Towards The Resilient Operator 5.0: The Future of Work in Smart Resilient Manufacturing Systems. In: *Procedia CIRP* 104, S. 1089–1094. DOI: 10.1016/j.procir.2021.11.183.

Rožanec, Jože M.; Lu, Jinzhi; Rupnik, Jan; Škrjanc, Maja; Mladenić, Dunja; Fortuna, Blaž et al. (2022): Actionable cognitive twins for decision making in manufacturing. In: *International Journal of Production Research* 60 (2), S. 452–478. DOI: 10.1080/00207543.2021.2002967.

Saboo, S.; Shekhawat, D. (2024): Enhancing Predictive Maintenance in an Oil & Gas Refinery Using IoT, AI & ML: An Generative AI Solution. In: Day 3 Wed, February 14, 2024. International Petroleum Technology Conference. Dhahran, Saudi Arabia, 12.02.2024 - 12.02.2024: IPTC.

Shin, Won; Han, Jeongyun; Rhee, Wonjong (2021): AI-assistance for predictive maintenance of renewable energy systems. In: *Energy* 221, S. 119775. DOI: 10.1016/j.energy.2021.119775.

Sorin, Vera; Brin, Danna; Barash, Yiftach; Konen, Eli; Charney, Alexander; Nadkarni, Girish; Klang, Eyal (2023): Large Language Models (LLMs) and Empathy – A Systematic Review.

Sun, Yicheng; Zhang, Qi; Bao, Jinsong; Lu, Yuqian; Liu, Shimin (2024): Empowering digital twins with large language models for global temporal feature learning. In: *Journal of Manufacturing Systems* 74, S. 83–99. DOI: 10.1016/j.jmsy.2024.02.015.

Touvron, Hugo; Martin, Louis; Stone, Kevin; Albert, Peter; Almahairi, Amjad; Babaei, Yasmine et al. (2023): Llama 2: Open Foundation and Fine-Tuned Chat Models. Online verfügbar unter http://arxiv.org/pdf/2307.09288v2.

Wang, Yuxia; Li, Haonan; Han, Xudong; Nakov, Preslav; Baldwin, Timothy (2023): Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs. Online verfügbar unter https://doi.org/10.48550/arXiv.2308.13387.

Wang, Zhuo; Bai, Xiaoliang; Zhang, Shusheng; Billinghurst, Mark; He, Weiping; Wang, Peng et al. (2022): A comprehensive review of augmented reality-based instruction in manual assembly, training and repair. In: *Robotics and Computer-Integrated Manufacturing* 78, S. 102407. DOI: 10.1016/j.rcim.2022.102407.

Wu, Qingyun; Bansal, Gagan; Zhang, Jieyu; Wu, Yiran; Li, Beibin; Zhu, Erkang et al. (2023): AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. Online verfügbar unter http://arxiv.org/pdf/2308.08155v2.

Xi, Zhiheng; Chen, Wenxiang; Guo, Xin; He, Wei; Ding, Yiwen; Hong, Boyang et al. (2023): The Rise and Potential of Large Language Model Based Agents: A Survey. Online verfügbar unter http://arxiv.org/pdf/2309.07864v3.

Yu, Hong Qing (2021): Dynamic Causality Knowledge Graph Generation for Supporting the Chatbot Healthcare

System. In: Kohei Arai, Supriya Kapoor und Rahul Bhatia (Hg.): Proceedings of the Future Technologies Conference (FTC) 2020, Volume 3, Bd. 1290. Cham: Springer International Publishing (Advances in Intelligent Systems and Computing), S. 30–45.

Yu, Wenhao (2022): Retrieval-augmented Generation across Heterogeneous Knowledge. In: Daphne Ippolito, Liunian Harold Li, Maria Leonor Pacheco, Danqi Chen und Nianwen Xue (Hg.): Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop. Hybrid: Seattle, Washington + Online. Stroudsburg, PA, USA: Association for Computational Linguistics, S. 52–58.

Zhao, Andrew; Huang, Daniel; Xu, Quentin; Lin, Matthieu; Liu, Yong-Jin; Huang, Gao (2023): ExpeL: LLM Agents Are Experiential Learners. Online verfügbar unter http://arxiv.org/pdf/2308.10144v2.

Zheng, Xiaochen; Lu, Jinzhi; Kiritsis, Dimitris (2022): The emergence of cognitive digital twin: vision, challenges and opportunities. In: *International Journal of Production Research* 60 (24), S. 7610–7632. DOI: 10.1080/00207543.2021.2014591.

Zhou, Bin; Li, Xinyu; Liu, Tianyuan; Xu, Kaizhou; Liu, Wei; Bao, Jinsong (2024): CausalKGPT: Industrial structure causal knowledge-enhanced large language model for cause analysis of quality problems in aerospace product manufacturing. In: *Advanced Engineering Informatics* 59, S. 102333. DOI: 10.1016/j.aei.2023.102333.

Zhu, Yinghao; Ren, Changyu; Xie, Shiyun; Liu, Shukai; Ji, Hangyuan; Wang, Zixiang et al. (2024): REALM: RAG-Driven Enhancement of Multimodal Electronic Health Records Analysis via Large Language Models. Online verfügbar unter http://arxiv.org/pdf/2402.07016v1.

Zigart, Tanja; Schlund, Sebastian (2020): Evaluation of Augmented Reality Technologies in Manufacturing – A Literature Review. In: Isabel L. Nunes (Hg.): Advances in Human Factors and Systems Interaction, Bd. 1207. Cham: Springer International Publishing (Advances in Intelligent Systems and Computing), S. 75–82.

BIOGRAPHIES

L. Kohl, Dipl.-Ing., has been a research assistant at the Institute of Management Sciences at TU Wien and at Fraunhofer Austria Research GmbH in factory planning and production management since September 2019. At Fraunhofer Austria, Linus Kohl leads the group for production optimization and maintenance management. Mr. Kohl studied industrial engineering - mechanical engineering at TU Wien. His work focuses on maintenance - the data-driven analysis and optimization of machines and systems using AI-based assistance systems. Linus Kohl is largely responsible for establishing the areas of retrofitting, maintenance as a service and industrial cognitive system at Fraunhofer Austria.

P. Besinger joined Fraunhofer Austria Research GmbH in 2021 in the role of research assistant, working on the development, integration and operation of AI-based software at companies and the associated requirements engineering. He completed his master’s degree in industrial engineering-Machine Engineering at TU Wien in 2021. Mr. Besinger received three TU Wien merit-based scholarships from 2018 - 2021 for outstanding performance. He is particularly interested in the application as well as implementation of Responsible AI in an industrial context to strengthen trust in AI-models and minimize socially harmful impacts of AI-systems.

S. Eschenbacher S. Eschenbacher is currently pursuing a Master’s degree in AI Engineering at the University of Applied Sciences Technikum Wien. She joined Fraunhofer Austria GmbH in 2022. Her research interest lies in the application of LLM and multi agent frameworks and on how these technologies can be deployed in real-world production environments to drive innovation and support the digital transformation in the industry.

F. Ansari is full professor at TU Wien and chair of Production and Maintenance Management and at the same time he serves as the head of strategic projects at Fraunhofer Austria. He conducts interdisciplinary research at the intersection of AI, Industrial Engineering and Production Management, where maintenance plays a central role. His interdisciplinary background is underlined by a degree in mechatronics and a dissertation in computer science (summa cum laude) at the University of Siegen. With his international involvement in various scientific associations (IEEE, IFAC, IALF), as well as his habilitation in Industrial Engineering, entitled "Management of Knowledge Intelligence in Human-centered Cyber Physical Production Systems", Dr. Ansari has established his role as an important part of the international research community.