JOINETECH

Application of Agentive Artificial Intelligence and Retrieval-Augmented Generation for Predictive Maintenance Operations

Salvador Perez-Garcia
Instituto Nacional de Estadística, Spain.
cossistas@hotmail.com

Citation:
Perez-Garcia, S. (2025). Application of Agentive Artificial Intelligence and Retrieval-Augmented Generation for Predictive Maintenance Operations. ***JOINETECH,*** 1(2), 187–191. https://doi.org/10.65479/joinetech.29

ARTICLE INFO

Keywords: Predictive maintenance, generative AI, AI agent, Industry 5.0, retrieval-augmented generation (RAG).

ABSTRACT

The advent of generative artificial intelligence (AI) for the general public and the fervor with which companies such as OpenAI, Google, Amazon, Microsoft, Meta, Alibaba, and High-Flyer (DeepSeek), among others, are competing to develop powerful solutions for generating text, images, audio, and video have revolutionized the methods available for obtaining and processing information. Within this context of technological advancement, this rapid development contrasts with the challenges that organizations face in harnessing this potential for the implementation of Industry 4.0 and its evolution toward Industry 5.0, with the aim of enhancing corporate performance and achieving significant differentiation from competitors. Furthermore, the impact on current organizational structures may be more profound than initially anticipated. As Tomlinson et al. (2025) outline, the jobs with the highest applicability for AI are notably those involving knowledge and communication, such as roles in computer science, mathematics, sales, and administration. The adoption of such solutions can lead to employee apprehension regarding their implementation, consequently necessitating the reorganization and redefinition of corporate hierarchies. In light of this situation, Anand and Wu (2025) recommend prioritizing these solutions at a strategic level, as opposed to a more traditional operational approach focused merely on performing existing tasks more rapidly. This strategic focus aims to create a differentiation that fosters the development of a sustainable competitive advantage. In this vein, the automation of tasks is increasingly highlighting the use of retrieval-augmented generation (RAG). When combined with large language models (LLMs), RAG facilitates collaboration and mutual learning with operators in administrative duties. This document presents the latest developments in the integration of generative and agentive AI for predictive maintenance tasks. However, such solutions are characterized by their requirement for specialized personnel for implementation and support, as well as the significant financial resources necessary for their deployment and ongoing maintenance.

Submission: October 9, 2025, Acceptance: November 24, 2025. Published: December 2025.

1. Introduction

When Industry 4.0 was introduced in the previous decade by the German government, it represented a significant impetus for enhancing the acquisition and management of manufacturing process data to support improved decision-making. Subsequently, the European Commission coined the term Industry 5.0, reinforcing industry's role as a contributor to society, rather than focusing solely on efficiency and productivity (European Commission, 2021).

The technological transformation in which contemporary organizations are immersed is characterized by the generation of massive amounts of data from autonomous sources. This is a direct consequence of the development of interconnectivity and networks, which necessitates appropriate data acquisition and storage. This, in turn, compels organizations to convert these vast quantities of data into actionable information and integrate this new knowledge into their structures (Wu et al., 2014).

This digital transformation entails significant changes in the organization through the integration of information, computing, communication, and connectivity technologies (Vial, 2021). This is closely aligned with Industry 5.0, as it is oriented toward the real-time analysis and processing of data and the intercommunication between various types of equipment and machinery, thereby enabling better decision-making support.

Furthermore, the technological revolution is advancing at a remarkable pace. It is currently characterized by the deployment of generative AI models, which are capable of creating novel data on the basis of learned patterns (Feuerriegel et al., 2024). This field is now evolving into agentive AI, which is designed to make decisions, solve problems, interact with its environment, and operate autonomously with the capacity to pursue complex goals with minimal supervision (Finn and Downie, 2025). This evolution is elevating large language models (LLMs), machine learning (ML), and natural language processing (NLP) to a new level.

This drive for digitalization is primarily oriented toward increasing productivity through enhanced responsiveness and adaptability to change. Significant academic interest has been sparked by these concepts, and this pressure is now being transmitted to organizations. In this context, organizations aim to maintain long-term commitment and investments (Deloitte, 2025). However, the benefits derived from the use of this technology stem mainly from intangible activities, such as improved customer relationships, the enhancement of existing products and services, or the shift of workers from performing low-level to high-level tasks, as opposed to the achievement of limited tangible benefits such as increased revenue, cost reduction, or productivity gains. This aligns with the observations of Ivanov et al. (2021), who highlight a lack of clarity regarding the realization of concrete economic benefits from the implementation of these technologies.

The effective application of this suite of measures within an organization is constrained, as a multitude of factors influence its performance. These include leadership style, the initial conditions of each company, size, target market, access to knowledge, the specific AI solution implemented, and the availability of adequate financial and human resources. The study by Oldemeyer, Jede, and Teuteberg (2025) identifies as many as 27 distinct challenges that organizations must overcome to successfully implement generative AI, among which economic, social, and technological dimensions are particularly prominent.

The deployment of this suite of technologies necessitates a highly qualified workforce capable of interpreting its outputs and implementing the corrections and enhancements required for proper functioning of AI systems. The most in-demand technical competencies include proficiency in programming languages, database management, machine learning, AI deployment, and AI security. Furthermore, soft skills are equally critical, encompassing communication and collaboration abilities, adaptability and continuous learning, critical thinking and problem-solving, and domain-specific knowledge (inuit.com, 2024). This underscores the significant level of upskilling an organization requires for an AI technician, coupled with the need for continuous learning capabilities to remain up to date in such a dynamic field.

Within a manufacturing enterprise, two distinct dimensions must be differentiated. The first pertains to administrative functions (human resources, marketing, accounting, procurement, etc.), while the second concerns operational aspects (supply chain, manufacturing, maintenance, etc.). In the administrative domain, AI agents and assistants have emerged with significant impact. These are primarily designed for the automation of operations, with some functioning reactively (assistants) and others proactively (agents), requiring minimal human supervision. Consequently, it is now common to find tools that streamline email composition, aid in real-time translation, enable the rapid identification of accounting discrepancies, or provide sales personnel with real-time information, thereby offering conversational support or even enabling the development of virtual sales representatives.

Within the operational domain, the feasibility of employing AI is highly contingent upon the quality of the data acquired and the investments necessary for its acquisition, processing, and storage. Generative AI can be utilized in various ways; for instance, computer numerical control (CNC) programming can be facilitated through the use of large language models that facilitate or collaborate in generating the necessary code. However, the most significant advantages are likely to be derived from maintenance tasks.

The digitization of technical documentation—typically provided in paper format by equipment manufacturers—when combined with augmented reality, enables operators to perform procedures in the correct sequence and validate completed operations in real time. This enhances standardization through improved reliability and procedural discipline. The next section outlines the characteristics of agentive AI in maintenance operations.

2. Retrieval-Augmented Generation

The limitations associated with the use of large language models (LLMs) primarily stem from their reliance on static pretraining data and their limited adaptability to dynamic environments. To address these constraints, retrieval-augmented generation (RAG) was introduced by Lewis et al. (2020). RAG integrates parametric and nonparametric memory for language generation. This architecture enables the LLM to access an external knowledge base—such as databases, documents, or application programming interfaces (APIs)—to ground its responses, thereby allowing LLMs to tailor their outputs to specific contexts (Belcic and Stryker, 2025). This approach mitigates issues such as model hallucination, the generation of responses with obsolete data, and the need for resource-intensive fine-tuning.

The principal distinction from a traditional automation system based on "if–then" loops is the capacity to analyze multiple variables simultaneously, incorporate the operational context, and evaluate the criticality of a situation before generating an action. In a maintenance environment, RAG combines advanced language models with internal databases (a knowledge layer) to enhance the accuracy and interpretation of fault detections. This synthesis empowers the agent to make or propose decisions on the basis of the specific situation and available knowledge.

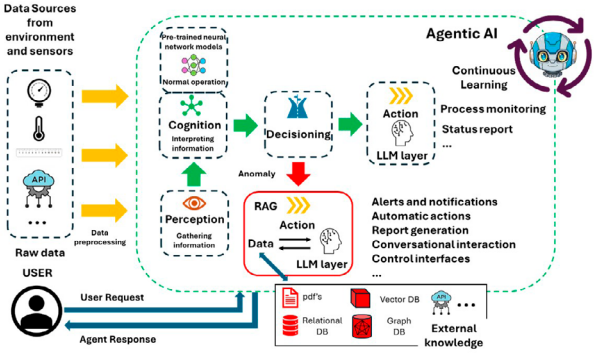

Fig. 1. ***The workflow of an AI agent utilizing RAG for predictive maintenance tasks (source: author).***

Figure 1 illustrates a workflow in which an AI agent captures environmental data and utilizes a set of neural networks pretrained on the equipment's normal operating conditions for each variable. In this framework, any anomalous data point is not properly interpreted by the network, resulting in a high interpretation error, which is then classified as a singular or irregular event. The application of neural networks in industrial equipment maintenance primarily focuses on machinery monitoring, optimization of electrical consumption, and quality control in manufacturing processes. These activities are fundamentally aligned with anomaly detection (AD) and predictive maintenance.

The literature on the use of such solutions is extensive. For instance, the types of neural network applicable to preventive maintenance tasks include recurrent neural networks (RNNs) such as long short-term memory (LSTM) networks (Zhou et al., 2020; Lin et al., 2020) and gated recurrent units (GRUs) (Azyus et al., 2023). Other architectures include one-dimensional (1D) or two-dimensional (2D) convolutional neural networks (CNNs), depending on whether the analysis is performed directly on the raw signal or on time–frequency representations, respectively (Guo et al., 2025; Apeiranthitis et al., 2024), as well as transformers adapted for time-series data (Perez-Garcia et al., 2025).

The security implications of this technology must also be addressed, requiring an assessment of risks to information confidentiality, integrity, and availability. Aligning with this concern, McKinsey (2025) enumerates potential security risks from such solutions, including the deployment of malicious agents to gain unauthorized privileges, agent spoofing to circumvent system safeguards, untraceable data exfiltration, and the spread of corrupted data that compromise agent decision-making.

It is essential to consider the legal framework with which this type of solution must comply. Within the European sphere, this is established by the General Data Protection Regulation (Regulation 2016/679, European Union, 2016), which protects individuals from purely automated decisions that affect them (Article 22). This implies that, while a vast number of processes can be automated, they must not generate decisions that directly impact individuals. On the other hand, the European Union (EU) AI Act (European Parliament, 2025) establishes that these generative AI solutions are not classified as high-risk, although it imposes the following limitations on their use: (1) indicating that the content was AI-generated, (2) designing the model to prevent the generation of illegal content, and (3) preventing the publication of copyright-protected data used in training.

The implementation of such solutions requires investments focused on the acquisition of licenses and subscriptions, the development and integration of the solutions, the provision of a computing infrastructure using graphics processing units (GPUs)/tensor processing units (TPUs), and user training. The benefits are not easily quantifiable in economic terms, such as improving problem resolution and decision-making, or relying on an expected increase in productivity. This renders their implementation a matter of trust rather than a purely economic decision (SiliconANGLE theCUBE, 2025)

3. Types of RAG Applicable in Maintenance

The RAG architectures employed depend primarily on the methods used for content retrieval and generation, characteristics that are intrinsically linked to performance, cost, and efficiency. The optimal solution is determined by the nature of the core components: the retriever (the information retrieval system), the augmentation engine (for context enhancement), and the generation engine (which produces a coherent, well-structured, and natural-language response). Accordingly, Harkar (2025) suggests that the most suitable configuration for industrial environments is advanced RAG, as it enhances both retrieval and generation through the use of sophisticated algorithms, re-rankers, fine-tuned LLMs, and feedback loops, compared with simpler solutions such as naive RAG or more complex ones such as modular RAG.

Furthermore, it is necessary to delineate the types of RAG on the basis of their primary objective. In this regard, two main categories are prominent: those focused primarily on improving reasoning with the acquired information, and those dedicated to enhancing the retrieval of relevant information through techniques such as query reformulation and the utilization of synonyms.

The first category is notably represented by reason and act (ReAct), which operates through a thought–act–observation cycle to address complex queries (Yao et al., 2022). This framework mitigates issues of hallucination and enhances reasoning by enabling the model to deconstruct a task, plan the subsequent logical step, and evaluate the outcomes. The model executes an action—typically a call to an external information source ("act"). This result is then fed back to the model as new information for a subsequent reasoning step ("observe"). To a lesser extent, chain-of-thought (CoT) prompting (Wei et al., 2022) can also be considered within this group. This technique decomposes a complex problem into sequential, intermediate reasoning steps before generating a final response.

The second category is prominently represented by query rewriting (Ma et al., 2023), which reformulates user queries to reduce the semantic gap between the input text and the knowledge required, thereby improving clarity and retrieval performance. To further enrich queries with synonyms or related terms, query expansion is employed (Jagerman et al., 2023). Figure 2 provides a potential use case for this technology, illustrating its application in the diagnosis and intervention by such agents to resolve an overheating issue in an industrial pump.

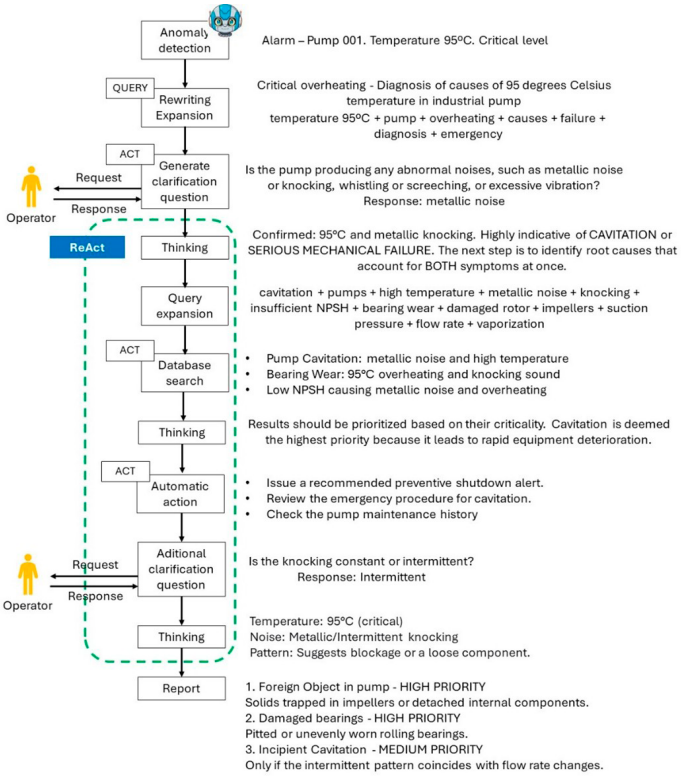

Fig. 2. ***An example workflow of a RAG system for anomaly detection (source: author).***

4. The Knowledge Base

The implementation of these solutions does not strictly require a relational database, where a list of problems is associated with a set of solutions. Instead, data lakes or vector databases can be utilized (Xinghan, 2025). An effective RAG system for maintenance applications requires access to a structured and unstructured repository of information in multiple formats and media. Its potential contents encompass maintenance personnel work orders, preventive maintenance standards (detailing time, parts, etc.), warehouse spare parts stock, purchase invoices, equipment manufacturer maintenance manuals, previous failure root cause analyses, anomalies and leaks identified in scheduled audits, internal regulations and standards, engineering specifications and internal documentation, and technical correspondence and supplier queries (Seneviratne et al., 2022; Wu et al., 2021). Therefore, the agent is intended to provide information on the potentially faulty component and the likely causes, along with an estimated repair time and cost, and the required resources.

5. Conclusions

This article has detailed the key challenges that companies face in deploying agentive AI for predictive maintenance. This technology requires high-quality data and its analysis by pretrained neural networks. Upon detecting an anomaly, the system automatically engages with personnel, retrieves technical and historical data, and executes protective actions.

The rapid digital shift from Industry 4.0 to 5.0 has placed real-time data interpretation at the heart of corporate strategy. Generative and agentive AI mark a qualitative leap in how manufacturers approach decisions, efficiency, and resilience. Yet, technology alone does not yield guaranteed economic returns. Success depends on leadership, resources, firm size, digital maturity, and the development of specialized skills and learning cultures.

AI finds strong applications in the maintenance field. Combining language models with retrieval systems (such as RAG) overcomes limitations of static data and factual inaccuracy in traditional LLMs. Use cases—from anomaly detection to augmented reality support—show clear potential to boost reliability and autonomy.

Ultimately, AI's effectiveness hinges on organized, high-quality knowledge bases that include manuals, logs, and real-time sensor data. For manufacturing, generative and agentive AI present a strategic opportunity—but one that demands long-term vision, investment in data governance, and cultural change. When embedded within a human-centered digital ecosystem, AI can become a cornerstone of industrial sustainability and resilience.

References

Anand, B. N., & Wu, A. (2025, November–December). *The Gen AI playbook for organizations: Where to use it, where not to, and why strategy still wins*. Harvard Business Review. https://hbr.org/2025/11/the-gen-ai-playbook-for-organizations

Apeiranthitis, S., Zacharia, P., Chatzopoulos, A., & Papoutsidakis, M. (2024). Predictive maintenance of machinery with rotating parts using convolutional neural networks. *Electronics, 13*, 460. https://doi.org/10.3390/electronics13020460

Azyus, A. F., Wijaya, S. K., & Naved, M. (2023). Prediction of remaining useful life using the CNN-GRU network: A study on maintenance management. *Software Impacts, 17*, 100535.

Belcic, I., & Stryker, C. (2025). *What is Agentic RAG?* IBM. https://www.ibm.com/es-es/think/topics/agentic-rag

Deloitte. (2025). *Now decides next: Generating a new future. Deloitte's State of Generative AI in the Enterprise. Quarter four report* (p. 19). https://www.deloitte.com/content/dam/assets-shared/docs/about/2025/quarter-4.pdf

European Commission. (2021). *Industry 5.0: Towards a sustainable, human-centric and resilient European industry*. https://doi.org/10.2777/308407

European Parliament (2025) *EU AI Act: first regulation on artificial intelligence* https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence#more-on-the-eus-digital-measures-3

European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119/1*. https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering, 66*(1), 111–126.

Finn, T., & Downie, A. (2025). *Agentive AI vs generative AI*. IBM. https://www.ibm.com/es-es/think/topics/agentic-ai-vs-generative-ai

Guo, H., Ping, D., Wang, L., Zhang, W., Wu, J., Ma, X., ... & Lu, Z. (2025). Fault diagnosis method of rolling bearing based on 1D multi-channel improved convolutional neural network in noisy environment. *Sensors, 25*(7), 2286.

Harkar, S. (2025). *RAG techniques*. IBM. https://www.ibm.com/es-es/think/topics/rag-techniques

Intuit Blog. (2024). *AI skills to boost your tech career*. https://www.intuit.com/blog/innovative-thinking/ai-skills-to-boost-your-tech-career/

Ivanov, D., Tang, C. S., Dolgui, A., Battini, D., & Das, A. (2021). Researchers' perspectives on Industry 4.0: Multi-disciplinary analysis and opportunities for operations management. *International Journal of Production Research, 59*(7), 2055–2078.

Jagerman, R., Zhuang, H., Qin, Z., Wang, X., & Bendersky, M. (2023). *Query expansion by prompting large language models* (arXiv:2305.03653). arXiv. https://arxiv.org/abs/2305.03653

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.

Lin, Z., Cheng, L., & Huang, G. (2020). Electricity consumption prediction based on LSTM with attention mechanism. *IEEJ Transactions on Electrical and Electronic Engineering, 15*. https://doi.org/10.1002/tee.23088

Ma, X., Gong, Y., He, P., Zhao, H., & Duan, N. (2023, December). Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 5303–5315).

McKinsey, (2025) *Deploying agentic AI with safety and security: A playbook for technology leaders*. McKinsey.com https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/deploying-agentic-ai-with-safety-and-security-a-playbook-for-technology-leaders#/

Oldemeyer, L., Jede, A., & Teuteberg, F. (2025). Investigation of artificial intelligence in SMEs: A systematic review of the state of the art and the main implementation challenges. *Management Review Quarterly, 75*(2), 1185–1227.

Perez-Garcia, S., Gonzalez-Gaya, C., & Sebastian, M. A. (2025). Enhancing asset reliability and sustainability: A comparative study of neural networks and ARIMAX in predictive maintenance. *Applied Sciences, 15*(10), 5266.

Seneviratne, D., et al. (2022). A Natural Language Processing (NLP) Framework for Asset Management Using Unstructured Data. *IEEE Transactions on Engineering Management.*

SiliconANGLE theCUBE (Oct., 2025) *Agentic AI ROI: From Automation to Decisions.* https://www.youtube.com/watch?v=PH6KNZykjX4&t=45s

Tomlinson, K., Jaffe, S., Wang, W., Counts, S., & Suri, S. (2025). *Working with AI: Measuring the applicability of generative AI to occupations* (arXiv:2507.07935). arXiv. https://arxiv.org/abs/2507.07935

Vial, G. (2021). Understanding digital transformation: A review and a research agenda. In *Managing digital transformation* (pp. 13–66).

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837.

Wu, X., Zhu, X., Wu, G., & Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering, 26*, 97–107.

Wu, C., et al. (2021). Intelligent knowledge graph construction for maintenance decision support in the aerospace industry. *Journal of Manufacturing Systems,* 60, 87-99.

Xinghan, P. (2025). *Comprehensive Guide to Choosing the Right Database for RAG Implementation: Leveraging Elasticsearch, Vector Databases, and Knowledge Graphs*. Medium.com. https://medium.com/@sampan090611/comprehensive-guide-to-choosing-the-right-database-for-rag-implementation-leveraging-47e7c6583fdc

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., & Cao, Y. (2022, October). ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Zhou, C., Fang, Z., Xu, X., Zhang, X., Ding, Y., & Jiang, X. (2020). Using long short-term memory networks to predict energy consumption of air-conditioning systems. *Sustainable Cities and Society, 55*, 102000.