



Vol. XVIII & Issue No. 06 June - 2025

INDUSTRIAL ENGINEERING JOURNAL

PROCON-AI: A RETRIEVAL-AUGMENTED LARGE LANGUAGE MODEL FOR INTELLIGENT QUERYING OF INDUSTRIAL PROCESS CONTROL SYSTEMS

Puneet Kaur

Associate Professor, Department of Electrical & Electronics Engineering, University Institute of Engineering & Technology, (UIET), Panjab University, Chandigarh
Email: puneetee@pu.ac.in / puneet2379@yahoo.com

Abstract

Industrial process control systems generate vast amounts of real-time sensor data, making it challenging to extract meaningful insights, particularly in relation to system dynamics and historical performance. This paper presents a Artificial Intelligence framework PROCON-AI, that leverages Large Language Models (LLMs) combined with Retrieval-Augmented Generation (RAG) using Lang Chain to enhance query capabilities for these systems. The proposed framework integrates real-time process data acquired from sensors, with domain-specific knowledge, enabling users to query operational status, fault diagnostics, and historical trends using natural language. The RAG mechanism enhances LLM responses by incorporating relevant information from industry-specific documentation and best practices, facilitating contextually rich and accurate outputs. The system is built using OpenAI's GPT-3.5 Turbo but is also compatible with open-source models like LLaMA2 and LLaMA3. Performance evaluations across multiple industrial processes show significant improvements in query accuracy, fault diagnosis speed, and response quality when compared to traditional SCADA systems and non-augmented LLMs.

1. INTRODUCTION

Industrial process control systems form the backbone of modern manufacturing, utilities, and chemical processing plants. These systems, ranging from water treatment facilities to pharmaceutical production lines, rely on complex networks of sensors, actuators, and control algorithms to maintain optimal operation. As these systems grow in complexity, the challenge of efficiently monitoring, troubleshooting, and optimizing them becomes increasingly daunting for human operators and engineers. Recent advancements in artificial intelligence, particularly in the domains of natural language processing and machine learning, offer promising solutions to these challenges. Large Language Models (LLMs) have demonstrated remarkable capabilities in understanding and generating human-like text across various domains. However, their application in highly specialized fields like industrial process control has been limited due to the need for up-to-date, domain-specific knowledge and real-time data integration.

The concept of Retrieval-Augmented Generation (RAG) [1] introduced by Lewis et al. represents a significant advancement in addressing knowledge-intensive NLP tasks. By combining pre-trained language models with a neural retriever that accesses external knowledge sources, RAG models demonstrate superior performance in tasks requiring specific factual information. This approach not only improves the accuracy and relevance of generated responses but also allows for more diverse and contextually appropriate outputs. The success of RAG in enhancing language models with external knowledge provides a promising framework for applications

that require both broad language understanding and access to specialized, up-to-date information.

In the domain of materials science, Chiang et al. introduced LLaMP [2], a multimodal retrieval-augmented generation framework that addresses the critical issue of LLM hallucinations in scientific applications. LLaMP employs multiple data-aware reasoning-and-acting agents to dynamically interact with computational and experimental data from the Materials Project. LaMP's success in handling diverse data types and reducing hallucinations in a highly specialized scientific domain illustrates the potential of RAG techniques for enhancing LLM performance in technical fields.

Lu et al. proposed a novel framework for self-improving additive manufacturing (AM) knowledge management [3], addressing the challenges in the evolving AM product development environment. Their four-tier framework integrates two key processes: a bottom-up data-driven knowledge engineering approach and a top-down goal-oriented active data generation method. This structure creates a closed loop system where AM knowledge can continuously evolve in an automated manner. By bridging the gap between accumulated data and practical knowledge application, this approach offers a pathway to more mature and efficient AM workflows, potentially revolutionizing how knowledge is managed and applied in complex manufacturing environments.

Rakholia et al. provide a comprehensive overview of the current state and future directions of Artificial Intelligence in the manufacturing sector [4]. Their review highlights how AI is driving intelligent automation across various aspects of

manufacturing, including predictive maintenance, quality control, process optimization, supply chain management, robotics, and decision support systems.

Sanchit et al. address the challenge of bridging the knowledge gap between Industry 4.0 technology developers and domain experts/end-users in the manufacturing sector. They propose an Industrial-GPT tool that leverages Large Language Models (LLMs) to translate natural language queries into meaningful inferences about asset-related data [5].

Haurum et al. showcase the application of Large Language Models (LLMs) in the real estate industry, demonstrating their potential as intelligent agents capable of engaging in open-ended conversations with investors. Their work focuses on developing a bilingual (English and Danish) real estate AI agent using LangChain and Pinecone [6], leveraging unstructured data from a professional Danish real estate website. The authors' approach addresses two significant challenges in real estate investment: the lack of professional knowledge among ordinary investors and the high costs associated with expert consultations. Through rigorous testing and evaluation, their AI agent demonstrated superior performance compared to other LLMs like Doubao and ChatGPT 4, producing more professional and concise outputs.

Arslan et al. address a critical challenge in the UK's transition to a net-zero economy by 2050, focusing on the role of Small and Medium-sized Enterprises (SMEs) in upgrading homes for energy efficiency. Their research introduces an innovative Energy Chatbot, a sustainable Information System that leverages Large Language Models (LLMs) integrated with multi-source Retrieval Augmented Generation (RAG) [7]. By consolidating diverse media sources, including news articles, government reports, industry publications, academic research, and social media, the Energy Chatbot provides SMEs with comprehensive, up-to-date information through a Question Answering (QA) system.

In their comprehensive literature review, Javaid, M and Singh, Ravi et al. explore the critical role of Artificial Intelligence (AI) in advancing Industry 4.0 [8]. The authors highlight AI's potential to enhance product consistency, boost productivity, and reduce operational costs in smart manufacturing environments. They emphasize how AI facilitates hyperconnected manufacturing processes by enabling machines to interact through automation systems that capture and interpret diverse data types.

Chandrasekhar et al. present AMGPT [9], a specialized Large Language Model (LLM) designed for contextual querying in additive manufacturing (AM). Their work addresses the limitations of generalized LLMs in providing specific, detailed information for materials science and manufacturing queries. Instead of training a new model from scratch, the authors leverage a pre-trained Llama2-7B model in a Retrieval-

Augmented Generation (RAG) setup, integrating information from approximately 50 AM papers and textbooks. Expert evaluations highlight that the RAG setup's specific embeddings improve response times and maintain coherence in generated text. This approach demonstrates the potential of specialized LLMs in providing detailed, context-specific information in technical fields like additive manufacturing, offering a valuable tool for researchers and practitioners in the field.

Melz introduces ARM-RAG (Auxiliary Rationale Memory for Retrieval Augmented Generation), an innovative approach to enhancing the intelligence of Large Language Models (LLMs) without the need for extensive retraining [10]. This method addresses a key limitation of frozen LLMs: their inability to learn from experience or acquire new knowledge over time. The ARM-RAG system builds upon the Retrieval Augmented Generation (RAG) framework by incorporating a mechanism to store and retrieve reasoning chains from successful problem-solving attempts. This work contributes to the ongoing efforts to make LLMs more dynamic and capable of continuous learning, potentially bridging the gap between artificial and human-like intelligence in problem-solving scenarios.

Jiang et al. proposes a LongRAG [11], an innovative framework that enhances retrieval-augmented generation (RAG) by leveraging long-context Large Language Models (LLMs). This approach addresses the imbalance in traditional RAG systems by processing Wikipedia into 4K-token units, significantly reducing the total number of units and easing the burden on the retriever. This work demonstrates the potential of combining long-context processing with RAG techniques, offering a new direction for improving AI systems in complex question-answering tasks and suggesting a promising future for integrating RAG with long-context LLMs in information retrieval and processing systems.

Drawing inspiration from these approaches, PROCON-AI discussed in this paper adapts and extends these techniques to the unique challenges of industrial process control. This system utilizes open-source LLM variants, including Llama2, Llama3, and Llama3.1, coupled with a custom-built RAG pipeline. This allows PROCON-AI to not only report current system states but also to provide contextual information, historical comparisons, and even predictive insights based on its understanding of the entire process control domain. Technique discussed in this paper, bridges the gap in applying Large Language Models (LLMs) to industrial process control by combining LLMs with Retrieval-Augmented Generation (RAG) techniques. The system is designed to provide an intuitive, natural language interface for querying complex industrial process control systems. By integrating real-time sensor data with a comprehensive knowledge base, including system documentation and domain-specific technical knowledge, it offers context-rich responses that go beyond simple data

retrieval. The solution delivers actionable insights by considering both the specific configuration of the process control plant and the current data from sensors, making it a valuable tool for operators and engineers to interact with and optimize these systems.

In section 2, architecture of PROCON-AI is discussed, along with the methodology for integrating diverse data sources, and present a comprehensive evaluation of the system's performance across various industrial processes in Section 3. The author has also discussed the potential limitations of AI-assisted process control. Finally, it concludes by outlining future research directions, including the potential for real-time adaptive learning and multi-modal data integration in industrial settings.

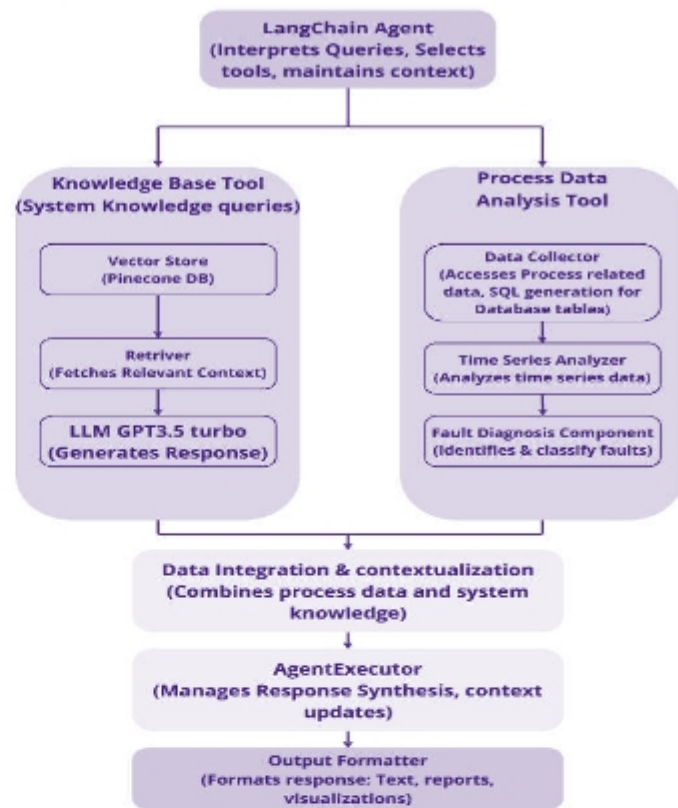
2. THEORY AND ARCHITECTURE

PROCON-AI's architecture is built with modularity and scalability in mind, relying on LangChain for natural language processing, knowledge retrieval, and real-time data analysis. The system begins with a user query interface, which accepts natural language inputs from users, allowing them to submit a wide range of queries. These queries may pertain to general system information, real-time operational statuses, or detailed fault diagnostics. The user interface seamlessly handles these inquiries, offering a flexible interaction mechanism. Fig.1 shows the High-level architecture of the solution. At the core of the architecture lies the LangChain Agent, which acts as the central orchestrator of query processing and response generation. This agent utilizes OpenAI's GPT-3.5 Turbo as the foundational language model. The LangChain Agent is responsible for interpreting the intent and context of user queries, dynamically selecting the appropriate tools required to generate an accurate response, and maintaining the conversation history for contextual continuity across multiple interactions.

The architecture integrates two key tools to manage knowledge retrieval and real-time data analysis. The Knowledge Base Tool is designed to handle queries related to system knowledge, utilizing a Retrieval-Augmented Generation (RAG) mechanism. It is underpinned by a vector store that uses a Pinecone database to store embeddings of system documentation and historical reports. The LangChain's Vector Store Retriever fetches the relevant context, which is then used by GPT-3.5 Turbo to generate contextually rich responses.

The Process Data Analysis Tool interacts with the process control system's real-time and historical databases. This tool is equipped with a data connector that leverages LangChain's SQL Database Chain to access relevant data. It includes a time series analyzer that performs statistical analysis and anomaly detection, while the fault diagnosis component, enhanced by machine learning models, identifies and classifies system faults based on predefined rules and patterns.

Fig.1: PROCON-AI, High-level architecture of the system



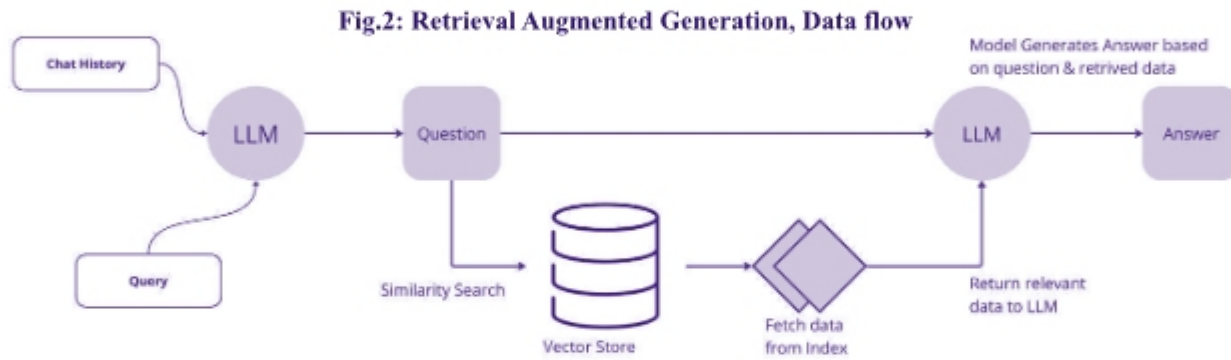
A crucial enhancement to the architecture is the Data Integration and Contextualization component. This component serves as a bridge between the Real-time Analysis Tool and the Knowledge Base Tool. When real-time data is retrieved and analyzed, it is passed through this component, where the LLM (GPT-3.5 Turbo) interprets the data in the context of the RAG-based system knowledge. This process ensures that real-time information is not presented in isolation but is enriched with relevant historical and contextual information from the system's knowledge base. The result is a more comprehensive and insightful analysis that combines current operational data with deep system understanding. Managing the overall execution flow is the Agent Executor component, which oversees tool execution, response synthesis, and context management. The Agent Executor receives the integrated and contextualized information from the Data Integration and Contextualization component. It then forms coherent and context-aware responses by combining this enriched data with any other relevant information from the query processing. LangChain's memory capabilities ensure that conversation context is maintained and updated throughout the process, enabling the system to handle complex, multi-query interactions.

The final system component is the Output Formatter, which structures the synthesized response for user presentation. This output may take the form of a simple text response for basic queries, structured reports for detailed analyses, or visualizations that represent real-time or historical system data, all enhanced with contextual insights. The entire PROCON-AI

architecture is built using LangChain's composable components, which facilitate easy modifications and extensions. The integration of RAG allows the system to incorporate new knowledge dynamically, without the need for retraining the underlying language model. The addition of the Data Integration and Contextualization component enables the system to provide more nuanced and informative responses, especially when dealing with complex industrial processes where context is crucial for proper interpretation of real-time data.

2.1 Retrieval Augmented Generation (RAG): Retrieval-Augmented Generation (RAG), introduced by Lewis et al. [1], represents a pivotal advancement in enhancing the capabilities

of LLMs by integrating dynamic, external knowledge bases. Retrieval Augmented Generation (RAG) represents a significant advancement in enhancing the capabilities of LLMs for domain-specific applications. RAG combines the generative power of LLMs with the ability to retrieve and incorporate external knowledge, addressing the limitations of traditional LLMs in specialized contexts. The core principle of RAG involves two main components. A Retriever, which identifies relevant information from a knowledge base in response to a query and another one, a Generator, which uses this retrieved information along with the query to produce a response. Fig.2 shows a typical information and query flow depicting a LLM creating response based on user query and supporting data fetched from a vector store.



In the context of industrial process control, RAG enables the system to leverage extensive technical documentation, operational manuals, and historical data while maintaining the flexibility and natural language understanding of LLMs.

Table 1: Components and Tools for Retrieval Augmented Generation based LLM

Function	Tool Name	Description
Preprocessing	<i>PyPDF2</i>	Tools for parsing and extracting text from PDF documents
Embedding Model	<i>GPT-3.5 Turbo with LangChain</i>	Utilizes GPT-3.5 Turbo integrated with LangChain for embedding and text generation tasks.
Vector Storage	<i>Pinecone or FAISS</i>	Vector databases for storing and managing embeddings generated by the embedding model.
Retrieval Agent	<i>LangChain's Retrieval Functions</i>	Manages retrieval of relevant information using LangChain's built-in retrieval capabilities.
Document Conversion	<i>Mathpix</i>	Converts PDF documents into TeX format to integrate into the Retrieval-Augmented Generation (RAG) pipeline.
Query Handling	<i>LangChain Agent</i>	Orchestrates query processing, tool selection, and maintains conversation context.
Response Generation	<i>GPT-3.5 Turbo</i>	Generates coherent and contextually accurate responses based on retrieved information.
Function	Tool Name	Description
Data Connector	<i>LangChain's SQLDatabaseChain</i>	Connects to real-time and historical databases for data retrieval and analysis.
Time Series Analysis	<i>Custom Statistical Tools</i>	Performs statistical analysis and anomaly detection on time series data
Output Formatting	<i>LangChain's Output Formatter</i>	Formats responses into text & structured data

2.2 Large Language Models (LLMs) in Industrial Applications:

Large Language Models (LLMs) have emerged as powerful tools in natural language processing, demonstrating significant capabilities in understanding and generating human-like text. However, their application in specialized domains like industrial process control presents unique challenges and opportunities. LLMs, such as GPT-3, Llama, and BERT, are trained on extensive general text data, enabling them to perform a wide range of language tasks. In industrial settings, these models offer the potential to interpret complex queries, generate detailed explanations, and assist in decision-making processes. However, the generalist nature of these models often leads to a lack of domain-specific knowledge crucial for industrial applications. The key challenge lies in bridging the gap between the broad knowledge of LLMs and the specialized expertise required in process control systems. This necessitates techniques for domain adaptation and knowledge integration, which form the cornerstone of the proposed system.

2.3 Process related Data Integration with System Knowledge:

A critical aspect of the system is the integration of time series process data with historical knowledge and static documentation. This integration enables the system to provide contextualized responses that consider both current operational states and established procedures. The main considerations in data integration include temporal alignment of data, removal of data inconsistencies and information prioritization. The temporal alignment of process data with historical records ensures insights are drawn from directly comparable operational contexts. Data inconsistencies and sensor errors are managed through validation mechanisms that verify the reliability of integrated information. A defined prioritization hierarchy weights current operational data above static information, while preserving the contextual value of established documentation and historical records. By combining runtime process data with a comprehensive knowledge base, the system can offer insights that are both timely and informed by historical trends and established best practices. This theoretical framework underpins the design and implementation of the AI-powered industrial process control system, addressing the unique challenges of applying advanced language models in specialized, critical industrial environments.

3. IMPLEMENTATION AND RESULTS

As highlighted in earlier sections, the primary challenge in industrial controls systems is the difficulty of extracting meaningful insights from vast, heterogeneous datasets generated by process control systems. Our approach focuses on solving practical problems faced by process engineers: how to quickly diagnose system faults, optimize performance, and interpret complex historical trends. To solve this, PROCON-AI integrates natural language interfaces with real-time data processing and retrieval-augmented generation (RAG) techniques.

3.1 Practical Problem Scenario: Water Treatment Plant: To test the model in a practical context, a mock water treatment plant was simulated. The simulated water treatment plant comprises key components such as intake pumps, screening and grit removal, chemical dosing systems, sedimentation tanks, filtration systems, disinfection units, clear water storage tanks, and distribution pumps. Various sensors monitor parameters like flow rates, pH levels, turbidity, chlorine levels, pressure, temperature, and pump status. The simulated system's process dataset was created in the database representing sensor readings from a water treatment plant. This data includes historical and

real-time values from sensors monitoring various components of the plant. To enhance the Retrieval-Augmented Generation (RAG) system, a comprehensive knowledge base was constructed using diverse documents relevant to water treatment. The knowledge base incorporated operational manuals containing detailed procedures for operating key components such as pumps, chemical dosing systems, and filtration units. These manuals provided essential operational insights, helping the AI model understand equipment functions and typical operating conditions. Datasheets of transmitters and sensors were included, providing specifications and calibration data for various sensors, including pH, chlorine, flow rate, and pressure transmitters. These datasheets enabled the system to interpret sensor readings accurately and provide guidance on normal versus abnormal values. The knowledge base also contained P and ID (Piping and Instrumentation Diagrams) representing the physical connections and control loops of the plant. These diagrams provided critical information on the relationships between components, helping the AI model understand how different units interacted and where specific sensors were located. Additionally, industry guidelines and standards documentation on best practices, safety standards, and water quality guidelines ensured the responses aligned with accepted norms and regulatory requirements.

3.2 Language Model and Retrieval-Augmented Generation (RAG): The core of this system is an advanced language model, OpenAI's GPT-3.5 Turbo, integrated with a RAG agent. When a query, such as

"What was the average pH level in the sedimentation tank yesterday?", is received, it follows a structured pipeline to analyze the input and arrive at final text output. The sequence of operation includes Query Analysis, Information retrieval and Context aware response generation steps.

3.2.1 Query Analysis and Embedding: The user query is embedded using the same vector representation approach used for the knowledge base.

3.2.2 Information Retrieval: The system retrieves the most relevant embeddings, such as recent sensor data logs and past incident reports related to pH levels.

3.2.3 Context-Enhanced Response Generation: The language model combines this retrieved information with the original query to generate an accurate response, including explanations of potential causes and corrective actions.

3.3 User Query and Results: System model was queried with different sentence and responses were captured. Few of such iterations are shown below.

Case Study 1:

User Input: "What was the average pH level in the sedimentation tank yesterday?"

Query Classifier Output: The query is classified as a database query for historical process data.

NL to SQL Conversion: The system converts the query to SQL: `SELECT AVG(ph_level)`

`FROM sensor_data`

`WHERE sensor_location = 'sedimentation_tank'`

`AND date = DATE_SUB(CURDATE(), INTERVAL 1 DAY);`

SQL Validator & Executor Output: The SQL query is validated and executed, yielding an average pH level of 7.2.

Result Analyzer Output: "The average pH level in the sedimentation tank yesterday was 7.2."

Contextualized Response Generator Output: *"The average pH level in the sedimentation tank yesterday was 7.2, which is within the optimal range of 6.8 to 7.5 for effective sedimentation. This indicates that the coagulation and flocculation processes are working effectively, promoting the formation of settleable flocs."*

Case Study 2:

User Input: *"Explain the purpose of coagulation in water treatment."*

System Knowledge Query Output: The language model provides a response based on its knowledge base: *"Coagulation is a critical step in water treatment that involves adding chemicals to neutralize charges on particles, causing them to form larger flocs, which can then be easily removed in subsequent processes."*

User Input: *"Show me the chlorine levels in the disinfection system for the past week."*

Query Classifier Output: The query is classified as a database query for historical data.

NL to SQL Conversion: The system converts the query to SQL:

```
SELECT timestamp, chlorine_level
FROM sensor_data
WHERE sensor_location = 'disinfection_system'
AND timestamp >= DATE_SUB(CURDATE(),
INTERVAL 7 DAY)
ORDER BY timestamp;
```

SQL Validator & Executor Output: The query is executed, and the results include a list of timestamps and chlorine levels.

Result Analyzer Output: *"Chlorine levels in the disinfection system over the past week ranged from 0.5 to 1.2 mg/L, with an average of 0.8 mg/L. There was a spike to 1.5 mg/L on Tuesday at 2 PM, which may require investigation."*

Contextualized Response Generator Output: *"Chlorine levels in the disinfection system over the past week ranged from 0.5 to 1.2 mg/L, with an average of 0.8 mg/L. This range is generally within acceptable limits for effective disinfection while minimizing the formation of disinfection by-products. However, there was a spike to 1.5 mg/L on Tuesday at 2 PM, which may require investigation. Possible causes for such a spike could include a malfunction in the chlorine dosing system or sudden changes in water quality."*

4. CONCLUSION AND FUTURE SCOPE

Based on the results discussed in the last section, the response quality of the system is significantly more insightful as compared to traditional database query results or statistical data report. Some of the key benefits of the PROCON-AI model-based process queries includes Enhanced Accessibility for users of process data metrics, context aware responses with comprehensive insights and improved decision making. By enabling natural language interactions, the system democratizes access to complex process data set, empowering both technical experts and non-technical stakeholders to derive actionable insights thus providing Enhanced Accessibility. Unlike standard BI tools, this solution incorporates intricate knowledge of system architectures, allowing for nuanced and context-rich analyses that go beyond simple data retrieval. Thereby exhibits Deep Contextual Understanding. The integration of detailed technical documentation and data-section mapping facilitates sophisticated analyses that capture the complex interplay between different fuel cell components, validating comprehensive insights. By providing architecture-informed insights, the system supports more effective operational strategies, predictive maintenance, and performance optimization. Future work may focus on expanding the system's capabilities to include predictive analytics, real-time monitoring, and integration with other process management

systems. As AI technologies continue to advance, there is also potential for incorporating more sophisticated machine learning models at different steps in the PROCON-AI data flow to further enhance the system's analytical capabilities. Currently GPT-3.5-turbo is used in the application but similar solutions can be developed by deploying LLMs like Llama, BERT, RoBERTa etc. These models have demonstrated excellent NLP capabilities and can be deployed to on-premises local hardware.

REFERENCES

- [1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks", in *Advances in Neural Information Processing Systems*, Vol. 33, Curran Associates, Inc., 2020, pp. 9459–9474.
- [2] Y. Chiang, C.-H. Chou, J. Riebesell, LLaMP: "Large language model made powerful for high-fidelity materials knowledge retrieval and distillation", in *arXiv: 2024, 2401.17244*
- [3] Y. Lu, Z. Yang, D. Eddy, S. Krishnamurty, "Self-Improving Additive Manufacturing Knowledge Management", in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Volume 1B: 38th Computers and Information in Engineering Conference, 2018, V01BT02A016*.
- [4] R. Rakholia, A. L. Suárez-Cetrulo, M. Singh and R. Simón Carbajo, "Advancing Manufacturing Through Artificial Intelligence: Current Landscape, Perspectives, Best Practices, Challenges, and Future Direction," in *IEEE Access*, vol. 12, pp. 131621-131637, 2024, doi: 10.1109/ACCESS.2024.3458830.
- [5] Sanchit, Bhattacharjee, Soumyabrata Pandhare, Vibhor, "Deriving inferences through natural language from structured datasets for asset lifecycle management", in *IFAC-Papers On Line*, 58, 145-150, 10.1016/j.ifacol.2024.08.064.
- [6] Kasper Raupach Haurum, Ruiqi Ma, Wen, "Long, Real Estate with AI: An agent based on LangChain", in *Procedia Computer Science*, Volume 242, 2024, Pages 1082-1088, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2024.08.199>.
- [7] Muhammad Arslan, Lamine Mahdjoubi, Saba Munawar, "Driving sustainable energy transitions with a multi-source RAG-LLM system", in *Energy and Buildings*, Volume 324, 2024, 114827, ISSN 0378-7788, <https://doi.org/10.1016/j.enbuild.2024.114827>.
- [8] Mohd Javaid, Abid Haleem, Ravi Pratap Singh, and Rajiv Suman, "Artificial Intelligence Applications for Industry 4.0: A Literature-Based Study", in *Journal of Industrial Integration and Management* 2022 07:01, 83-111
- [9] Achuth Chandrasekhar, Jonathan Chan, Francis Ogoke, Olabode Ajenifujah, Amir Barati Farimani, "AMGPT: A large language model for contextual querying in additive manufacturing", in *Additive Manufacturing Letters*, Volume 11, 2024, 100232, ISSN 2772-3690, <https://doi.org/10.1016/j.addlet.2024.100232>.
- [10] Melz, Eric, "Enhancing llm intelligence with arm-rag: Auxiliary rationale memory for retrieval augmented generation.", in *arXiv preprint arXiv:2311.04177 (2023)*
- [11] Jiang, Ziyang, Xueguang Ma, and Wenhui Chen, "Longrag: Enhancing retrieval-augmented generation with long-context llms.", in *arXiv preprint arXiv:2406.15319 (2024)*