جامعة ابن طفيل
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵙⵉ ⵎⵓⵃⴰⵎⵎⴷ
Ibn Tofaïl University
Faculty of Sciences

Ibn Tofail University
Faculty of Sciences, Kénitra

Final Year Project Report

Master Of Artificial Intelligence And Virtual Reality

Reverse Osmosis Performance Optimization: A Comprehensive Approach Integrating Artificial Intelligence and LLM for Malfunction Prediction and Real-Time Troubleshooting in Reverse Osmosis Systems

---

Host establishment : **O N E E K é n i t r a**

Prepared by : Mr. Abdellah WALID & Miss. Ihssane AOUNE

Supervised by : Mr. Younes TOUMINE (ONEE Kenitra)
Dr. Abdelhakim EL FADIL (Aquadviser)
Pr. Khaoula BOUKIR (ENSC Kenitra (UIT))

D e f e n d e d o n : ***September 19th, 2024, before the jury composed of :***

- Pr. Raja TOUAHNI (FS Kenitra (UIT))
- Pr. Anass NOURI (FS Kenitra (UIT))
- Pr. Rochdi MESSOUSSI (FS Kenitra (UIT))
- Pr. Khaoula BOUKIR (ENSC Kenitra (UIT))
- Pr. Souad Eddarouich (CRMEF Rabat)
- Dr. Abdelhakim EL FADIL (Aquadviser)

Academic Year 2023/2024

Acknowledgments

First and foremost, we would like to express our gratitude to God. To our family, and to our friends for their unwavering support and encouragement throughout this journey. Their constant presence and belief in us have been a source of strength and motivation.

We are profoundly grateful to Pr. Khaoula Boukir, who honored us by serving as our academic supervisor. Her unwavering support, guidance, and understanding provided the moral strength and inspiration needed to bring this project to fruition.

We would also like to extend our sincere thanks to Dr. Abdelhakim El Fadil and Dr. Aboubakr Achraf El Ghazi, Founder and Co-Founder of Aquadviser. We deeply appreciate the opportunity they provided us to collaborate with the Aquadviser team, which greatly enriched our learning experience.

We are also grateful to the entire teaching faculty of the AIVR Master's program. In particular, we would like to thank Pr. Raja Touahni for her assistance and support throughout this journey.

Our heartfelt thanks extend to Mr. Younes Toumine, Head of the Methods Office at ONEE and our external supervisor. His invaluable advice and constant encouragement played a crucial role in the successful completion of this work. We also wish to extend our deep appreciation to Mr. Ismael Chaoui, Head of the TAG Kenitra Central at ONEE Kenitra, for his support and guidance throughout this project.

Additionally, we express our sincere gratitude to all members of the jury, who have graciously agreed to evaluate our work. We hope that this project meets their expectations and stands as a testament to the quality of education we have received.

Lastly, with the greatest respect, we extend our heartfelt thanks to all the teachers and administrative staff of Ibn Tofail University for their continued dedication and support.

Table of Contents

List of Figures

List of Tables

Introduction

Recent advancements in artificial intelligence have pushed it to the forefront of scientific and technological discourse. AI's impact surpasses disciplinary boundaries, influencing fields as diverse as healthcare diagnostics, algorithmic finance, and industrial automation. Given AI's widespread influence, understanding its potential for environmental sustainability is important.

One specific area of promise lies in reverse osmosis (RO) , that is a membrane based process technology to purify water by separating the dissolved solids from feed stream resulting in permeate and reject stream for a wide range of applications in domestic as well as industrial applications. It is seen from literature review that RO technology is used to remove dissolved solids, colour, organic contaminants, and nitrate from feed stream [1].

Reverse osmosis systems offer a valuable solution for water treatment, particularly in areas with limited freshwater resources. However, these systems face several challenges that hinder their efficiency and sustainability. Major problems include membrane fouling, which reduces water flow and increases energy consumption. Fouling causes significant loss of productivity and added operational cost thus becomes a challenge on membrane operation [2]. Additionally, RO processes can generate a significant amount of brine waste, posing environmental concerns. Energy consumption per unit of water produced is also a crucial performance measure, as it directly impacts the environmental footprint of the RO system.

This thesis has two main contributions, each addressing a critical aspect of improving the management of reverse osmosis systems.

The first one focuses on how machine learning can predict problems in reverse osmosis (RO) systems before they happen. We will use algorithms and methods to learn from past information, like water quality, previous issues, and flow rates, to spot patterns that might lead to things going wrong. By knowing problems are coming, we can fix them early and avoid expensive shutdowns. By implementing these algorithms, Professionals in RO systems can save money, keep things running smoothly, and make water treatment more sustainable.

The second contribution consists of using Large Language Models (LLMs) to perform in the troubleshooting of problems related to RO systems. This section will explore how LLMs can be integrated into existing maintenance frameworks to support engineers and technicians in

identifying root causes and implementing solutions more quickly and accurately.

The findings presented here aim to contribute to the ongoing efforts to optimize water treatment processes, offering practical insights that can be applied across various industrial contexts.

The structure of this report begins with Chapter 1, which outlines the context of the internship, including the motivation, objectives, and project planning. Chapter 2 delves into the theoretical foundations of reverse osmosis (RO) systems, addressing major challenges like biofouling and scaling, followed by a review of AI techniques for predictive maintenance and the role of Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG). Chapter 3 focuses on the practical application of machine learning for predicting water quality and RO system performance, including data pre-processing and model evaluation. Chapter 4 transitions to the implementation of LLMs in troubleshooting RO systems, supported by experimental results and case studies. The report ends with offering a synthesis of findings and recommendations for future research and practical applications in RO performance optimization.

Chapter 1: Context of the internship

1.1 Team

This chapter provides an overview of the context in which the internship was conducted, highlighting the collaborative environment and key teams involved.

To better understand the dynamics of the internship, it is essential to introduce the three main teams that played a pivotal role in shaping this experience: Team Ibn Tofail, Team Aquadviser, and Team ONEE. Each of these teams brings a unique set of skills and expertise to the table, contributing to the success of the Project. The following sections will provide detailed descriptions of each team, their core functions, and their contributions to the projects we were involved in during the internship.

1.1.1 Team Ibn Tofail

Team Ibn Tofail is composed of two Master's students, Ihssane Aoune and Abdellah Walid, specializing in Artificial Intelligence. Both students are enrolled in the Master's program in Artificial Intelligence and Virtual Reality at the Faculty of Sciences in Kenitra. Both students are under the academic supervision of Pr Khaoula Boukir at Ibn Tofail University.

To introduce new insights into the water desalination field, Team Ibn Tofail collaborates closely with professionals who specialize in water treatment and desalination technologies. By combining their expertise in AI with the industry knowledge of these professionals, the team aims to develop AI-driven solutions that can optimize desalination processes.

1.1.2 Team Aquadviser

Aquadviser is a company dedicated to advancing water desalination technologies by integrating artificial intelligence into troubleshooting processes. Founded by Abdelhakim El Fadil, who

leads the development of Aquadviser's services and forms strategic partnerships. Co-founder Aboubakr Achraf El Ghazi, who manages technical developments and oversees operations, plays a crucial role in driving the company's technical excellence and operational efficiency.

Aquadviser's team collaborates closely with Team Ibn Tofail, providing assistance with technical problems and offering critical evaluation of the AI team's results. This collaboration ensures that the AI-driven solutions developed by Team Ibn Tofail are effectively created to meet the practical needs of the desalination industry. Moreover, Aquadviser plays an important role in facilitating communication with Team ONEE, coordinating efforts to align the AI developments with the operational requirements and standards of ONEE. By combining advanced AI expertise with practical industry experience, this partnership aims to introduce new insights and innovations, enhancing the reliability and efficiency of water desalination through cutting-edge technology.

1.1.3 Team ONEE

Team ONEE (National Office of Electricity and Drinking Water), is focused on ensuring efficient water supply and distribution across Morocco. The team is composed of Younes Toumine who is the external supervisor for our internship, and Head of the Methods Office within the company, and Ismael Chaoui, Head of the TAG Kenitra Central. Both professionals helped us gain a deeper understanding of how systems operate within the company. Their insights were valuable in bridging the gap between AI technologies and practical applications in desalination.

ONEE contributed significantly to the project by providing essential datasets from their operations, which were crucial for developing and testing our AI-driven solutions. Their support went beyond data provision; they also ensured that our solutions were aligned with the real-world needs of the desalination industry.

1.2 Motivation and Objectives

1.2.1 Introduction

The complexity of managing reverse osmosis (RO) systems has long posed challenges in the water treatment industry. Ensuring the consistent performance of these systems is essential to maintaining water quality standards. RO systems are essential for producing increased-quality water but are prone to failures that can compromise their efficiency and reliability.

1.2.2 Problem Identification

Despite their importance in delivering high-quality water, RO systems face issues such as membrane fouling and scaling, which can compromise their efficiency and reliability. Therefore, ensuring good performance is essential for meeting water quality standards and minimizing operational costs. Additionally, these failures result in a waste of time, resources, and money. Thus, it will be crucial to develop and implement better solutions that can accurately predict and effectively address system problems to help minimize the negative impact and ensure more efficient operations.

For that purpose, Aquadviser focused on applying artificial intelligence to solve water treatment problems. They first noticed a frequent issue: despite the advanced technology used in RO systems, unpredictable failures frequently occurred, leading to increased operational costs and resource waste.

1.2.3 Motivation

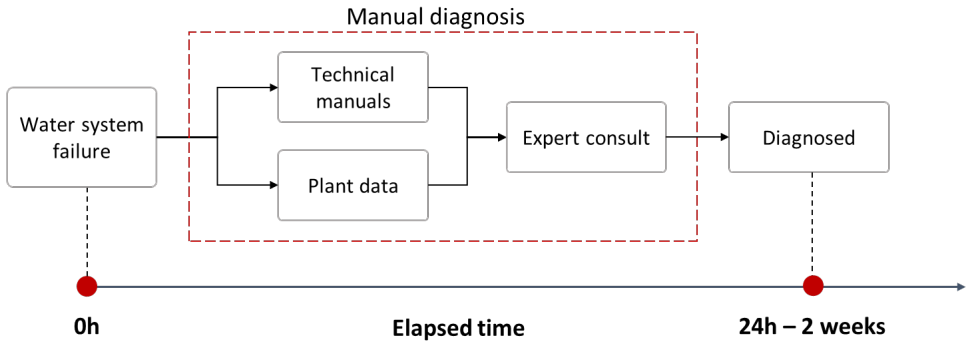

Figure 1.1: A timeline illustrating how Inefficient troubleshooting amplifies water treatment costs

Mid-size desalination operations experience an annual loss of approximately €1.5 million due to sub-optimal performance [3] [4]. This significant financial impact highlights the pressing need to optimize the troubleshooting process, aiming to reduce costs and minimize the time required for problem resolution. By focusing on improving efficiency and effectiveness in troubleshooting, these operations can achieve better economic outcomes and enhance overall system performance.

Recognizing the potential to improve the reliability and efficiency of these systems through the use of AI, Aquadviser initiated a collaboration with ONEE (National Office of Water and Electricity), a Moroccan company that manages the electricity and water supply. It focuses on rural electrification, monitoring electricity tariffs, and executing various projects to ensure

reliable service in Morocco. ONEE helped provide datasets from different RO systems, that contain multiple parameters influencing the performance of RO systems and failure incidents. These datasets were used to develop advanced predictive models.

With the help of both companies, our collective goal was to develop innovative solutions that not only predict potential issues within RO systems before they lead to significant failures but also streamline the troubleshooting process when problems do occur.

This collaboration has led to the development of a two-part approach. The first involves using machine learning models to analyze the characteristic data of RO systems (Using ONEE datasets), enabling the early prediction of system failures. The second focuses on utilizing Large Language Models (Using technical manuals and data from reliable online sources) to assist in resolving issues, providing real-time, AI-driven insights to support maintenance teams.

Together, Aquadviser, ONEE, and our research team worked to transform the maintenance and operation of reverse osmosis systems, aiming to lower maintenance costs and improve overall system performance. The solutions being developed in this collaboration are expected to offer a proactive and intelligent approach to managing RO systems.

1.2.4 Objectives

The main objectives of our project are developed in the figure below :

Develop Predictive Models for RO Systems

Our team focuses here on predicting potential failures and performance issues in reverse osmosis (RO) systems using machine learning techniques. By analyzing datasets provided by ONEE, which include various parameters influencing the performance of RO systems, we aim to develop predictive models that can identify potential failures before they occur. This task required careful data cleaning and preprocessing to ensure the accuracy and reliability of the models. By anticipating issues, the model helps in maintaining system efficiency and preventing costly breakdowns, thereby enhancing the overall reliability of water treatment processes.

Enhance Troubleshooting with AI-Driven Insights

To enhance the troubleshooting experience, we implement Large Language Models (LLMs) to offer real-time guidance for maintenance teams dealing with RO systems. With the help and expertise of Aquadviser's team, we extracted valuable data from reliable online sources, such as technical manuals and research publications. This data, combined with the field knowledge of industry experts, allows the LLMs to provide accurate and context-specific insights for

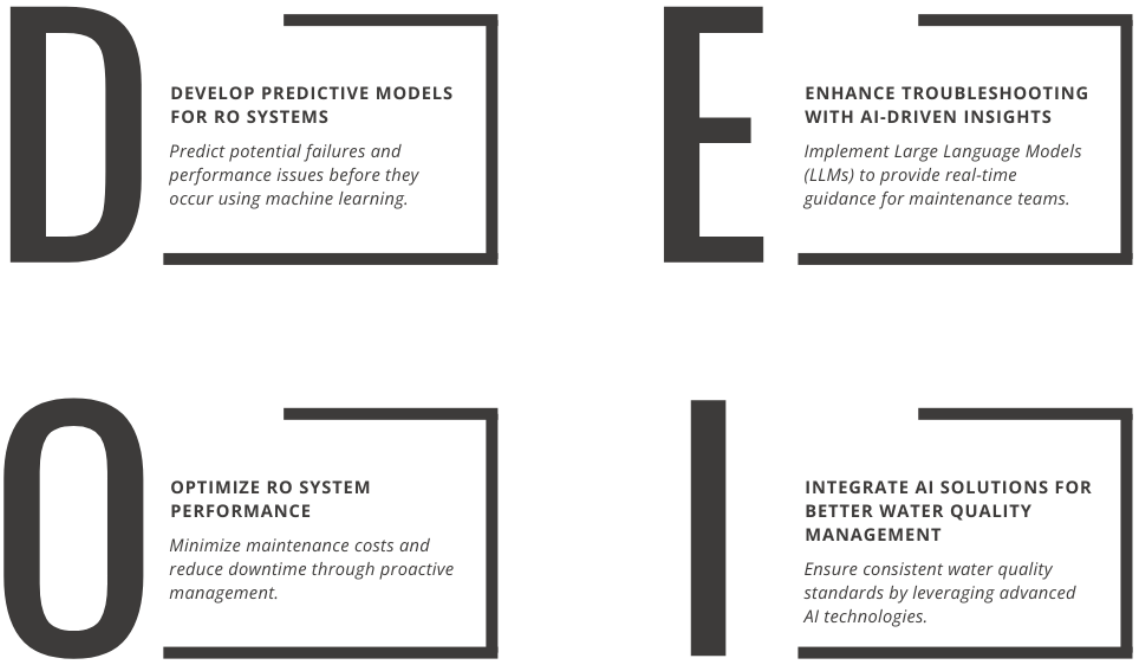

Figure 1.2: Main Objectives of the Thesis

troubleshooting. By facilitating quick and informed decision-making, these AI-driven insights improve the effectiveness of maintenance efforts and reduce the time required to address system issues.

Optimize RO System Performance

By using LLMs to streamline the troubleshooting process, professionals can save significant time and resources in diagnosing and resolving issues. The optimized management of RO systems ensures continuous operation with minimal disruptions, ultimately leading to cost savings and improved performance.

Integrate AI Solutions for Better Water Quality Management

Using feedback from industry experts, the AI solutions are created to provide high-quality guidance and support to maintenance teams. By integrating AI, the water treatment industry can achieve more precise control over water quality, ensuring that output consistently meets regulatory standards. The integration of AI-driven insights and predictive models forms a comprehensive approach to maintaining and improving water quality management practices.

1.3 Planning

1.3.1 Gantt

A Gantt chart, commonly used in project management, is one of the most popular and useful ways of showing activities (tasks or events) displayed against time. On the left of the chart is a list of the activities and along the top is a suitable time scale. Each activity is represented by a bar; the position and length of the bar reflects the start date, duration and end date of the activity [5].

1.3.2 Tasks

Figure 1.3: Timeline of the project using Gantt

As illustrated in Figure 1.3, the tasks are described in greater detail below:

Literature Review

The initial phase involves a literature review to understand the current state of reverse osmosis (RO) systems, AI, and machine learning applications in water treatment. The findings from this

phase will guide the development of predictive models and AI-driven solutions.

Data Collection and Processing

This phase focuses on collecting and preprocessing data from ONEE, which includes parameters related to the performance of RO systems and failure incidents. The data is cleaned and processed to ensure accuracy and reliability, to build strong predictive models. This phase is critical as the quality of data directly impacts the effectiveness of the machine learning models.

Building Prediction Models

Once the data is prepared, the next phase involves developing predictive models using machine learning techniques. These models aim to forecast potential failures and performance issues in RO systems based on historical data.

Data Scraping of Documents

Data scraping is used to extract documents from reliable online sources. This data is used to inform the development of Large Language Models (LLMs), which provide real-time troubleshooting guidance for maintenance teams. Aquadviser's expertise in the field is leveraged during this phase to ensure the accuracy and relevance of the extracted data.

RaG on Q&A File and Evaluation

In this phase, Retrieval-Augmented Generation (RaG) techniques are applied to the Q&A file that contains data from Dupont's Q&A website, which was collected to generate insightful responses. The generated answers are evaluated for accuracy and usefulness, and the model is refined accordingly. This iterative process aims to improve the model's ability to provide precise and context-aware guidance.

GraphRAG Research and Implementation

Research and implementation of GraphRAG techniques are undertaken to incorporate graph-based data into AI models. This phase explores how GraphRAG's Microsoft can enhance the quality of the generated answers.

GraphRag vs. RaG Evaluation

This task involves comparing the performance of GraphRAG with traditional RaG methods in generating AI-driven insights for troubleshooting. The evaluation will determine which approach provides better accuracy and reliability in managing RO systems.

Aquadviser Web App

This final phase involves developing and integrating a user-friendly web app for Aquadviser that includes a chatbot interface. The web app is designed to provide maintenance teams with real-time access to AI-driven insights, improving the efficiency of troubleshooting and decision-making processes. The web app also includes a homepage page that contains details about Aquadviser.

Chapter 2: Theoretical Background

2.1 Introduction

This chapter outlines the main theoretical concepts and technologies relevant to this research. It starts with an overview of Reverse Osmosis (RO) systems and their challenges in water treatment. The chapter then discusses the use of AI for Predictive Maintenance on datasets provided by ONEE, utilizing machine learning techniques to enhance system reliability by predicting potential failures. It also covers Large Language Models (LLMs) for Troubleshooting, which provide AI-driven insights to maintenance teams dealing with RO systems. Additionally, the chapter introduces Retrieval-Augmented Generation (RaG) and GraphRAG techniques, which were used to retrieve information from technical manuals and data scraped from reliable sources, therefore improving information retrieval and understanding in complex systems. These theories and methods collectively support the development of innovative solutions for better managing RO systems.

2.2 Reverse Osmosis

2.2.1 Principles of Reverse Osmosis

Reverse osmosis (RO) is a widely used water purification technology that uses a semi-permeable membrane to remove ions, unwanted molecules, and larger particles from drinking water. The process was developed in the mid-20th century to desalinate seawater and has since become the dominant method for desalination due to its efficiency and effectiveness in producing fresh water. RO is employed extensively for seawater desalination, wastewater treatment, and water purification in industrial and municipal settings.

At its core, reverse osmosis relies on applying pressure to force water molecules across a semi-permeable membrane. The membrane selectively allows water molecules to pass through while retaining dissolved salts and other impurities. In this process, water flows from a region of

higher solute concentration (i.e., saltwater or brackish water) to a region of lower concentration, effectively separating fresh water from the solute.

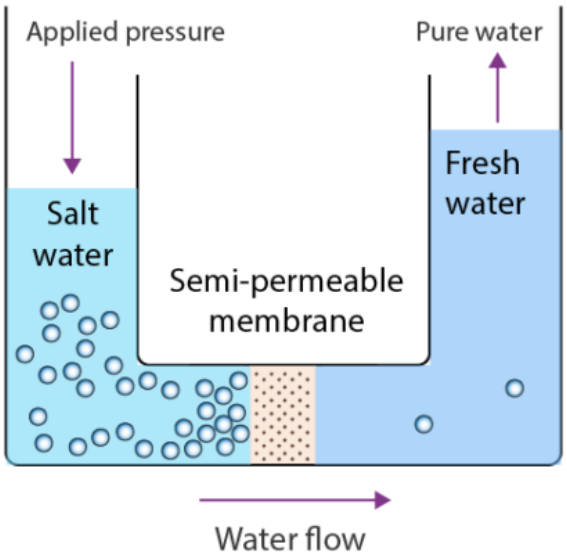

Figure 2.1: Diagram of the Reverse Osmosis Process

The driving force behind reverse osmosis is the application of pressure that exceeds the natural osmotic pressure of the water, which would normally cause water to move in the opposite direction (from low to high concentration). By overcoming this natural tendency, RO systems can produce highly purified water, with recovery rates typically ranging from 40% to 50%, depending on the system's design and the water source.

2.2.2 Historical Context and Development

The development of RO as a practical water treatment technology began in the 1950s and 1960s. Researchers initially focused on creating membranes that could withstand the pressures necessary for desalination while maintaining selectivity for water molecules. Early RO membranes were made from cellulose acetate, but modern advancements led to the creation of more durable and efficient thin-film composite membranes.

The technology gained significant traction in the following decades, especially as water scarcity became a growing concern globally. Modern RO systems have evolved to become more energy-efficient and reliable, largely driven by improvements in membrane technology and system design [6]. The development of energy recovery devices has been instrumental in reducing the overall energy consumption of RO plants, making them a more viable option for large-scale desalination.

2.2.3 Applications of Reverse Osmosis

The most prominent application of reverse osmosis is in seawater desalination. In countries with limited freshwater resources but access to the ocean, RO has become a crucial technology for ensuring a sustainable water supply. Large-scale seawater desalination plants, particularly those using seawater reverse osmosis (SWRO), are now operational in countries like Spain, as well as in many Gulf countries [7].

Beyond desalination, RO is also used for brackish water treatment, industrial wastewater recycling, and water purification in various sectors. For instance, brackish water desalination can augment water supplies in inland regions, although it faces challenges related to the management of brine disposal [7].

2.2.4 Major Problems in Reverse Osmosis Systems

Reverse osmosis (RO) systems face several challenges that can significantly impact their performance and efficiency. The two most prevalent issues are scaling and biofouling [8].

Scaling

Scaling occurs when dissolved minerals in the feed water precipitate on the membrane surface or within the membrane pores. Common scaling compounds include calcium carbonate, calcium sulfate, barium sulfate, and silicates. Scaling clogs the membrane pores, reducing the permeate flux and increasing the pressure required to maintain the desired flow rate. If left unchecked, scaling can lead to complete membrane failure [8].

Biofouling

Biofouling is the accumulation of microorganisms, such as bacteria, algae, and fungi, on the membrane surface. These organisms form a biofilm that traps other particles, further exacerbating the fouling problem [8] [9]. Biofouling increases the trans-membrane pressure, reduces permeate flux, and can degrade product water quality [8]. It is considered the most complex and challenging type of fouling in RO systems [9].

Other types of fouling, such as colloidal fouling and organic fouling, can also occur in RO systems. Colloidal fouling involves the deposition of suspended particles, while organic fouling is caused by the accumulation of organic matter on the membrane [10].

To mitigate these issues, RO systems employ various pretreatment methods, such as filtration, disinfection, and chemical treatment [8] [9]. Effective pretreatment is crucial for reducing the

fouling and scaling potential of the feed water [8]. Regular membrane cleaning and replacement are also necessary to maintain the system's performance and extend the membrane's lifespan [10].

2.2.5 Challenges and Sustainability

Despite its widespread adoption, reverse osmosis faces several challenges. Energy consumption remains a significant concern, as seawater desalination requires substantial energy input to pressurize the water. According to Menachem Elimelech and William A. Phillip, current RO technology consumes between 3 and 4 kWh per cubic meter of water produced, which is still more energy-intensive than traditional freshwater treatment methods. The energy consumption is compounded by environmental concerns associated with brine discharge and potential harm to marine ecosystems due to the high salinity and chemical contaminants in the wastewater [7].

There is a growing emphasis on improving the sustainability of RO through innovations such as advanced membranes, enhanced energy recovery, and the use of renewable energy sources. New research is focused on the development of fouling-resistant membranes that can reduce the need for chemical cleaning and pretreatment, further lowering operational costs and minimizing environmental impact [7].

2.3 Artificial Intelligence Techniques for Predictive Maintenance

2.3.1 Introduction

In recent years, the integration of artificial intelligence (AI) and machine learning (ML) techniques in water treatment has gained significant attention, particularly for reverse osmosis (RO) systems. These systems are crucial for desalination and water purification, yet they often face challenges such as membrane fouling and performance degradation. The application of AI offers promising solutions to predict water quality and identify potential problems early, enabling more efficient management and maintenance of RO systems. This section explores the use of machine learning for quality water prediction and problem identification, reviewing key studies that show the effectiveness of these techniques in enhancing RO system performance.

Here are the main goals of applying AI for predictive maintenance :

- **AI for Quality Water Prediction** The goal here is to apply AI for water quality prediction in RO systems to enhance the accuracy and efficiency of predicting key performance indicators, such as salt passage, permeate flow rate, and pressure difference. By using

machine learning models to analyze various input variables, AI can provide precise predictions that help optimize water treatment processes and ensure consistent water quality.

- **AI for Problem Identification**

  The second goal is to use AI for the identification of potential problems in RO systems. By employing data-driven models, we will be able to reduce operational costs, and extend the lifespan of the RO membranes, thereby enhancing overall system reliability.

2.3.2 Algorithms used for Quality Water Prediction

In a study on factors affecting reverse osmosis membrane performance [11], the researchers employed various machine-learning techniques to predict the output variables and analyze the impact of different factors on RO membrane performance

In the context of machine learning for quality water prediction, models like ANNs, Random Forests, and MLR play a critical role. Random Forest models excel in predicting salt passage, using variables such as temperature and conductivity to assess membrane performance. Permeate flow rate, a key indicator of water production efficiency, is effectively predicted by MLR, which analyzes factors like feed flow rate and membrane characteristics. For pressure difference predictions, ANNs are highly effective due to their ability to handle complex, non-linear relationships between operational variables. These predictions are crucial for optimizing water treatment processes and ensuring consistent water quality in RO systems.

2.3.3 Algorithms used in predictive maintenance

In a study on data-driven model approach related to the application of AI in the predictions [12], Several AI and machine learning algorithms have proven effective for predictive maintenance in reverse osmosis (RO) systems. Artificial Neural Networks (ANNs) are widely used for their ability to model complex relationships between input variables and system performance, making them effective for predicting non-linear factors like pressure differences. Random Forest is another popular algorithm, known for its accuracy and robustness in large datasets, particularly when predicting salt passage. Support Vector Machines (SVMs), known for their high accuracy, are often used to simulate changes in flow rate and conductivity. Multiple Linear Regression (MLR), though simpler than the others, is efficient for predicting variables like permeate flow rate, especially when linear relationships exist between variables. Lastly, Long Short-Term Memory (LSTM) networks, a form of deep learning, are used to predict time-dependent variables like transmembrane pressure (TMP), allowing for accurate predictions of membrane fouling using historical data.

2.4 Large Language Models (LLMs)

2.4.1 Introduction

Large Language Models (LLMs) have become essential tools in the field of artificial intelligence, particularly in natural language processing. Over time, these models have evolved from simple rule-based systems to complex architectures capable of understanding and generating human-like text. This section provides an overview of the key developments and technologies that have shaped LLMs, highlighting their growing importance and the challenges they present in modern AI applications.

2.4.2 Historical Background

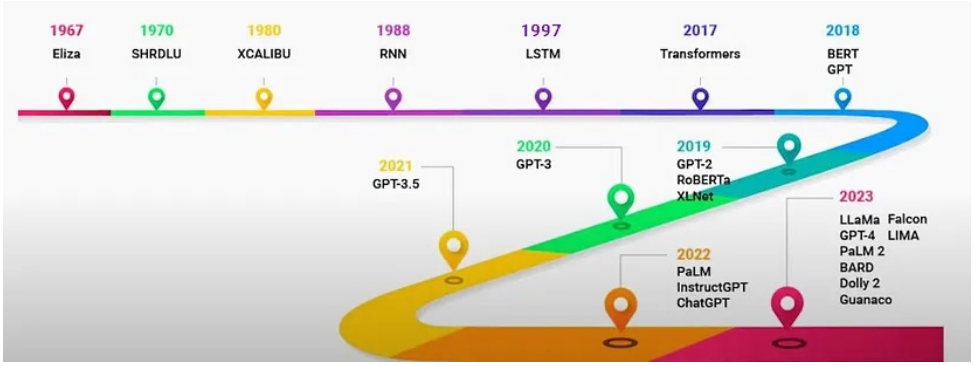

Figure 2.2: Large language model (LLM) timeline.

It is necessary to explain that Large Language Models (LLMs) have experienced an incredible transformation during the past few decades and have become the backbone of natural language processing (NLP) and the overall artificial intelligence (AI) domain. The journey of LLMs began in the 1950s and 1960s with early rule-based systems and initial attempts at machine translation, such as the Georgetown-IBM experiment in 1954 [13], which demonstrated the potential of computational language processing. In 1966, Joseph Weizenbaum's creation of ELIZA [14], the first chatbot, marked a significant milestone in simulating human conversation through pattern recognition techniques.

The 1970s to 1990s witnessed a shift towards statistical models in NLP, with the adoption of techniques like Hidden Markov Models (HMMs) and n-gram models, which utilized statistical co-occurrences to predict word sequences. This period also saw the introduction of neural networks in language modeling, most notably with the development of the Long Short-Term Memory (LSTM) architecture in 1997 [15], addressing challenges like the vanishing gradient problem in recurrent neural networks (RNNs).

The deep learning revolution of the 2000s and 2010s significantly advanced LLMs. Researchers like Yoshua Bengio pioneered the use of feed-forward neural networks for language modeling in 2001, laying the foundation for deep learning in NLP. The introduction of Word2Vec in 2013 [16] allowed for the representation of words as continuous vectors in high-dimensional space, enhancing the capture of semantic relationships. However, the most transformative development came with the Transformer architecture in 2017 [17], which replaced RNNs with self-attention mechanisms, enabling parallel processing and greatly improving training efficiency.

Since 2018, LLMs have rapidly evolved with the advent of pre-trained language models like BERT and GPT-2, demonstrating the power of larger models trained on extensive datasets. Notable models such as GPT-3, with 175 billion parameters, showcased the ability to perform a wide array of tasks with remarkable fluency. This evolution continues with recent innovations like GPT-4o and LLaMA3, which further push the boundaries of what LLMs can achieve, setting the stage for even more advanced applications and research in the field.

2.4.3 The Evolution of Modern LLMs

The Transformer Architecture: The Building Block

The transformer architecture is the fundamental building block of all Large Language Models (LLMs). The transformer architecture was introduced in the paper “Attention is all you need,” published in December 2017 [17].

There are seven important components in transformer architecture. that work together to process and generate text:

1. **Inputs and Input Embeddings:** The tokens entered by the user are transformed into numerical representations called “input embeddings.” that the model can process. These embeddings represent words as vectors in a mathematical space, enabling the model to understand and compare word meanings based on their proximity.

2. **Positional Encoding:** Unlike traditional neural networks, transformers incorporate positional encoding to ensure that the model understands the sequence in which words appear, which is crucial for maintaining the correct meaning in sentences.

3. **Encoder:** The encoder first tokenizes the input text into a sequence of tokens, such as individual words or sub-words. It then applies a series of self-attention layers to generate hidden states that encapsulate the meaning and context of the text at various abstraction levels. This information is crucial for the next steps in the model.

4. **Outputs (shifted right):** During training, the decoder learns how to guess the next word by using the preceding words, with the sequence shifted to the right.

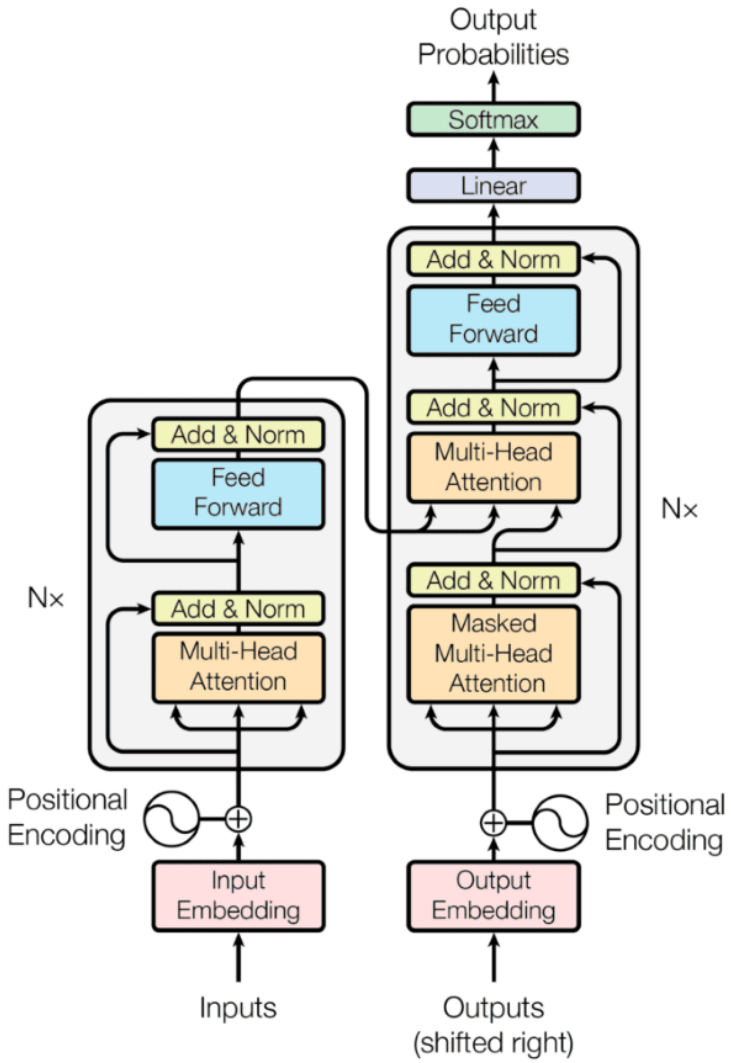

Figure 2.3: Transformer architecture.

5. **Output Embeddings:** the output must be changed to a numerical format, known as output embeddings. Output embeddings are similar to input embeddings and go through positional encoding, which helps the model understand the order of words in a sentence. A loss function is used in machine learning, which measures the difference between a model's predictions and the actual target values.

6. **Decoder:** The positionally encoded input representation and the positionally encoded output embeddings go through the decoder. The decoder is part of the model that generates the output sequence based on the encoded input sequence. During training, the decoder learns how to guess the next word by looking at the words before it.

7. **Linear Layer and Softmax:** After the decoder produces the output embeddings, the linear layer maps them to a higher-dimensional space. This step is necessary to transform the output embeddings into the original input space. Then, we use the softmax function to generate a probability distribution for each output token in the vocabulary, enabling us to generate output tokens with probabilities.

A significant innovation in the transformer architecture is the Attention Mechanism. Unlike previous models like RNNs and LSTMs, which process inputs sequentially, transformers can analyze the entire input sequence simultaneously. This attention mechanism allows the model to focus on different parts of the input sequence selectively, enabling it to capture long-term dependencies and relationships between words more effectively. This capability makes transformers particularly powerful in handling complex language tasks.

LLaMa3: Advancements in Large Language Modeling

LLaMA 3 (Large Language Model Meta AI) is part of the LLaMA family of language models, including LLaMA and LLaMA 2. LLaMA 3 represents a significant leap in the capabilities of open-source AI models, designed to compete with some of the most powerful proprietary models [18].

LLaMA 3 introduces several innovations in its architecture and training process. It utilizes a decoder-only transformer architecture and incorporates an expanded tokenizer with a vocabulary of 128,000 tokens, which enhances its ability to process and generate text across multiple languages. This version also features Grouped-Query Attention (GQA) to improve inference efficiency and scalability, particularly for handling longer context windows, making it more adept at tasks like summarization and complex reasoning [19].

The training of LLaMA 3 involved a massive dataset of over 15 trillion tokens, significantly larger than that used for LLaMA 2, with a strong emphasis on high-quality multilingual and domain-specific data, including coding and reasoning tasks. This extensive dataset, combined with advanced training techniques such as data, model, and pipeline parallelization across thousands of GPUs, allowed LLaMA 3 to achieve state-of-the-art performance on various benchmarks, and outperform some of the most powerful proprietary models [18].

GPT-4o: Pushing the Boundaries of LLM Capabilities

On May 13, 2024, OpenAI introduced GPT-4o, its latest flagship large language model (LLM). The “o” in GPT-4o stands for “Omni,” which signifies the model’s ability to handle multiple modalities, such as text, audio, and images, all within a single model. This multimodal approach enables GPT-4o to engage in real-time conversations, answer questions, generate text, and process various forms of media more efficiently than its predecessors, including GPT-4 Turbo [20].

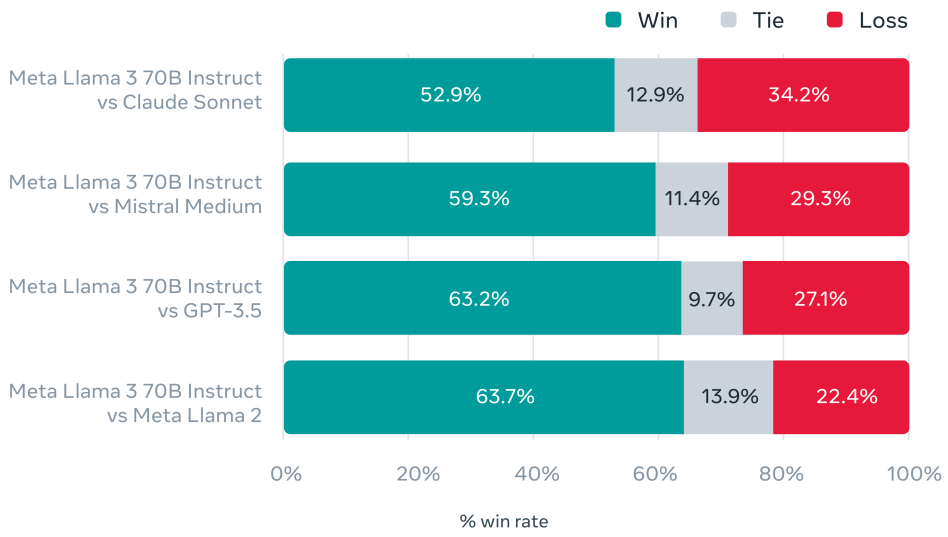

Figure 2.4: LLaMa3 Human evaluation (aggregated).

GPT-4o represents a major advancement in generating high-quality text efficiently. One of its key strengths is the ability to generate text at twice the speed of its predecessor, GPT-4 Turbo, while reducing costs by 50

The model supports a context window of 128,000 tokens, enabling it to manage extensive and complex text inputs and outputs. This large context window is particularly valuable for tasks that require sustained, coherent text generation across multiple steps, such as long-form content creation, detailed reports, or intricate dialogue systems. Moreover, GPT-4o can produce up to 4,096 tokens per request, allowing it to generate substantial amounts of text in a single output [21].

2.4.4 Challenges in Implementing LLMs

The implementation of Large Language Models (LLMs) presents several significant challenges:

- **Non-Deterministic Responses:** LLMs can produce variable outputs for the same input, which complicates their reliability in scenarios requiring consistent responses, such as in medical or legal contexts. Developers need to implement robust validation mechanisms to ensure accuracy.
- **Observability and Monitoring:** The complexity of LLM workflows necessitates comprehensive monitoring to evaluate the quality of outputs and detect errors. Tools for observability can help trace issues across LLM applications, allowing teams to analyze performance and security risks effectively.

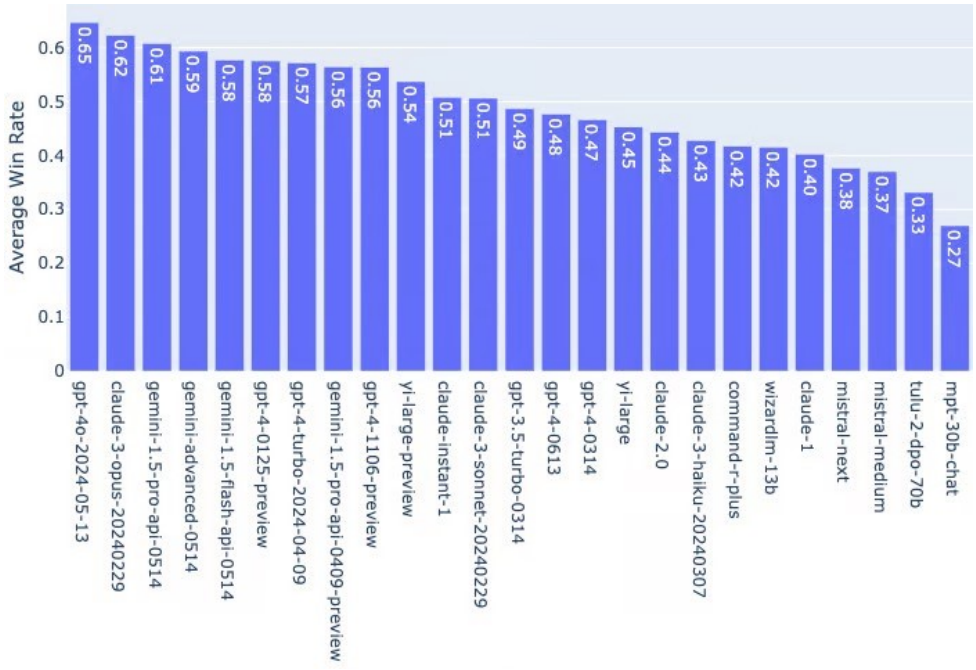

Figure 2.5: GPT-4o Human evaluation.

- **Scalability and Resource Management:** LLMs require significant computational resources, which can lead to challenges in scaling applications to meet user demand. Efficient resource management is crucial for maintaining performance and cost-effectiveness.
- **Security Risks:** LLMs can be vulnerable to security exploits, such as prompt injection attacks, which may lead to data exposure or inappropriate outputs. Implementing security measures is essential to safeguard against these vulnerabilities [22].

2.4.5 Conclusion

The evolution of Large Language Models has been marked by continuous innovation and significant breakthroughs, from the early days of rule-based systems to the sophisticated, multimodal models of today. As demonstrated by the advancements in LLaMA3 and GPT-4o, LLMs have reached new heights in their ability to process and generate language across diverse contexts and modalities. However, the implementation of these models is not without challenges, such as ensuring consistency, managing computational resources, and safeguarding against security risks. As we transition to the next section on Retrieval-Augmented Generation (RAG), it becomes clear that the future of LLMs will likely involve hybrid approaches that combine the strengths of LLMs with other AI techniques to address these challenges and further enhance the capabilities of language-based AI systems.

2.5 Retrieval-Augmented Generation (RaG)

2.5.1 Introduction

In recent years, the field of natural language processing (NLP) has seen remarkable advancements, particularly with the development of large language models (LLMs) that have revolutionized tasks such as text generation, translation, and question-answering. However, these models, while powerful, often struggle with factual accuracy, sometimes producing plausible-sounding but incorrect or nonsensical responses—a phenomenon known as "hallucination." To address these challenges, researchers have introduced Retrieval-Augmented Generation (RAG), a sophisticated architecture that enhances the relevance and accuracy of generated responses by integrating retrieval mechanisms with LLMs.

At its core, RAG leverages a dual approach merging traditional search functionalities with the dynamic prompting capabilities of LLMs. This hybrid model enables the system to ground its responses in retrieved data, providing a robust foundation for generating more accurate answers.

2.5.2 Historical Background

The concept of RAG builds upon the evolution of information retrieval and natural language generation techniques. Some Early efforts in the 2000s to combine retrieval-based methods with generative models aimed primarily at improving information retrieval systems by integrating simple generative models for response synthesis. However, the true potential of RAG was realized with the advent of large-scale pre-trained language models in the late 2010s, such as BERT and GPT. These models provided a powerful foundation for natural language understanding and generation, but they often struggled with factual accuracy, leading to the integration of retrieval mechanisms.

Formalizing RAG as an architecture was introduced by researchers at Facebook AI in 2020 [23], who demonstrated that combining a dense retrieval mechanism with generative transformers significantly improved the quality of the generated text. By grounding it in retrieved documents. This advancement marked a major step forward in the development of AI systems, enabling them to produce more truthful and contextually relevant responses.

2.5.3 How RAG Works

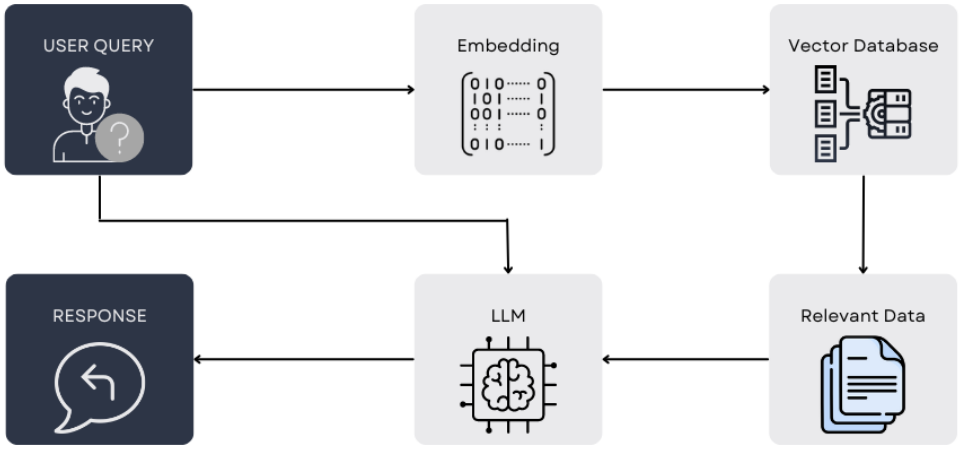

Figure 2.6: Workflow of RAG.

Without RAG, the LLM takes the user input and creates a response based on the information it was trained. RAG introduces an information retrieval component that first utilizes user input to pull information from a new data source. The user query and the relevant information are both given to the LLM. The LLM uses the new knowledge and training data to create better responses.

The new data outside of the LLM's original training dataset is called external data. It can come from multiple data sources, such as APIs, databases, or document repositories. These data sources can be in various formats, including files, database records, or long-form text. A technique called "chunking" is employed to manage large and complex datasets effectively. Chunking involves breaking down the data into smaller, more manageable pieces, or "chunks," which are then processed individually. This allows the system to handle vast amounts of information efficiently. Each chunk is processed using embedding language models, which convert the information into numerical representations and store it in a vector database, creating a knowledge library that the generative AI models can access and understand.

Once the external data is chunked and embedded, the system performs a relevancy search by converting the user query into a vector representation and matching it against the entries in the vector database. For example, consider a smart chatbot that can answer human resource questions for an organization, if an employee asks, "How much annual leave do I have?", the system will retrieve annual leave policy documents alongside the individual employee's past leave record. These specific documents will be returned because they are highly relevant to what the employee has input. The relevancy was calculated and established using mathematical vector calculations, often using "cosine similarity".

Next, the RAG model augments the user input (or prompts) by adding the relevant retrieved data

in context. This step uses prompt engineering techniques to communicate effectively with the LLM. The augmented prompt allows the large language models to generate an accurate answer to user queries [24].

2.5.4 Advantages of RAG

RAG offers several advantages:

- **Improved Accuracy:** By accessing up-to-date external data, RAG helps mitigate the inaccuracies arising from LLMs static training data [25].
- **Reducing inaccurate responses, or hallucinations:** Integrating retrieved information reduces the likelihood of generating incorrect or fabricated information (hallucinations) [26]. Trust and Transparency: By allowing models to cite sources, RAG enhances user trust in the generated content, providing a mechanism for users to verify the claims made by the AI [27].
- **Flexibility and Efficiency:** RAG allows organizations to customize LLM outputs to specific domains or contexts without the high costs and time associated with retraining models. This adaptability is particularly valuable for customer support and knowledge management systems applications [24].
- **Being efficient and cost-effective:** Compared to other approaches to customizing LLMs with domain-specific data, RAG is simple and cost-effective. Organizations can deploy RAG without needing to customize the model. This is especially beneficial when models need to be updated frequently with new data [28].

2.5.5 Applications of RAG

RAG has found applications across various domains, demonstrating its versatility and effectiveness:

- **Customer Support:** Automated chatbots powered by RAG can provide accurate answers to customer inquiries by retrieving relevant information from knowledge bases or FAQs.
- **Content Creation:** RAG can assist writers and content creators by generating articles or reports that are grounded in the latest research and data, ensuring accuracy and relevance [25].

- **Educational Tools:** RAG can enhance learning platforms by providing students with contextually rich explanations and answers, drawing from a wide array of educational resources.
- **Research Assistance:** RAG systems can aid researchers in quickly finding relevant literature and synthesizing information, streamlining the research process.

2.5.6 Challenges and Future Directions

Despite its advantages, the RAG architecture is not without challenges. Key issues include:

- **Retrieval Quality:** The effectiveness of RAG relies heavily on the quality of the retrieved documents. Poor retrieval can lead to inaccurate or irrelevant responses.
- **Computational Complexity:** The combined processes of retrieval and generation can be computationally intensive, necessitating efficient algorithms and infrastructure.
- **Bias and Fairness:** Like all AI systems, RAG models can inherit biases present in the training data or retrieved documents, raising concerns about fairness and ethical implications.

Future research in RAG may focus on improving retrieval algorithms, enhancing the integration of diverse data sources, and addressing ethical considerations to ensure responsible AI deployment. One promising direction is the development of GraphRAG, an extension of the RAG framework that explores new ways to leverage graph-based data structures in the retrieval process, promising further advancements in the field.

2.6 GraphRAG

2.6.1 Introduction

The paper titled “From Local to Global: A Graph RAG Approach to Query-Focused Summarization” [29] presents a new method for improving the capabilities of large language models (LLMs) in summarizing large datasets. This method, Graph RAG, aims to address the limitations of existing retrieval-augmented generation (RAG) techniques, particularly when dealing with global questions that require insights from text.

2.6.2 Motivation

The motivation behind the development of Graph RAG originates from the limitations of traditional retrieval-augmented generation (RAG) methods when addressing global questions that require comprehensive insights from large datasets. While RAG effectively retrieves localized information to answer specific queries, it struggles with general, query-focused summarization tasks, such as identifying themes within texts. This challenge is particularly noticeable when dealing with large volumes of data, where the context window of large language models (LLMs) may not encompass all relevant information. Consequently, the authors propose a more robust solution that combines the strengths of RAG and query-focused summarization (QFS) by utilizing a graph-based indexing system. This system allows for the efficient organization of information into communities of related entities, enabling the generation of coherent and comprehensive summaries that can effectively respond to global queries. By leveraging community detection algorithms within graphs, Graph RAG improves the ability to summarize and synthesize information from big datasets, making it a powerful advancement over traditional RAG methods [29].

2.6.3 Overview of Graph RAG

Key Concepts

- **Retrieval-Augmented Generation (RAG):** A technique that combines retrieval of relevant information from external sources with generative capabilities of LLMs to answer questions or summarize texts.
- **Query-Focused Summarization (QFS):** A task that involves generating summaries that are specifically tailored to answer a given query, rather than just extracting relevant text.

The authors argue that traditional RAG methods struggle with global questions (e.g., “What are the main themes in the dataset?”) because they typically focus on localized text retrieval. In contrast, the Graph RAG approach integrates a graph-based indexing system that allows for more comprehensive summarization.

Methodology

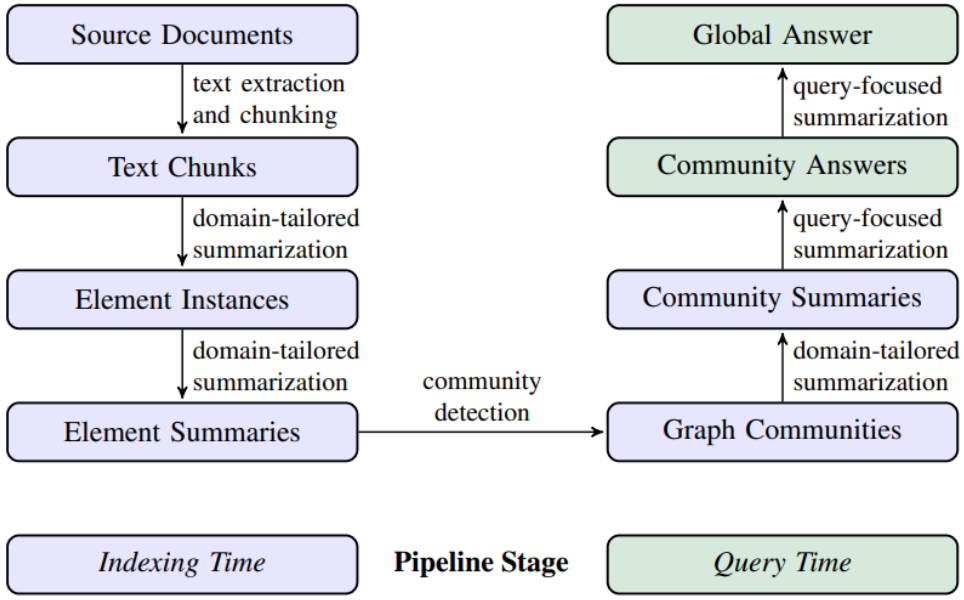

Figure 2.7: Graph RAG pipeline using an LLM-derived graph index of source document text.

Graph-Based Indexing

1. **Entity Knowledge Graph Creation:** The first step involves extracting entities and their relationships from the source documents to create a graph index. This is done using LLMs which analyze text chunks to identify entities, relationships, and relevant claims.
2. **Community Detection:** The graph is then split into communities of closely related nodes using community detection algorithms (Leiden algorithm). This method allows the model to summarize related entities together, improving the relevance and coherence of the generated summaries.
3. **Community Summarization:** Each community is summarized independently, creating a set of community summaries that can be used to answer user queries.
4. **Answer Generation:** When a user poses a question, the system generates partial answers from the community summaries. These partial answers are then combined and summarized to produce a final, comprehensive response.

Evaluation and Results

The authors evaluated the Graph RAG approach using two datasets, each containing around one million tokens. They compared the performance of their method against traditional RAG and other summarization techniques. The results indicated that Graph RAG significantly outperformed naive RAG in terms of the comprehensiveness and diversity of the generated answers [29].

The method effectively handled global sensemaking questions, demonstrating its efficiency in summarizing large volumes of text [29].

The findings suggest that Graph RAG can facilitate better understanding and insights from large datasets, making it a valuable tool for tasks that require comprehensive analysis, such as scientific research, intelligence analysis, and educational purposes. In summary, this approach not only addresses the limitations of existing methods but also opens doors for future research in improved sense-making and information retrieval.

Chapter 3: Predicting water quality and predictive maintenance using machine learning

3.1 Introduction

In this section, we explore the use of machine learning techniques for predicting water quality and implementing predictive maintenance in reverse osmosis (RO) systems. Ensuring high water quality and system reliability are crucial for the efficient operation of RO systems, which are widely used in water treatment processes. In this study, we used datasets provided by ONEE Kenitra and ONEE Khenifra to make our prediction. By leveraging data-driven models, we aim to anticipate changes in water quality and detect potential system failures before they occur, allowing for proactive maintenance and reducing downtime. This part covers the entire process, from the motivation and objectives behind this approach to data collection and pre-processing, model development, results, and evaluation. Additionally, we discuss the limitations of our current models and suggest directions for future research to further enhance predictive capabilities in water quality management and system maintenance.

3.2 Predicting water quality in RO systems

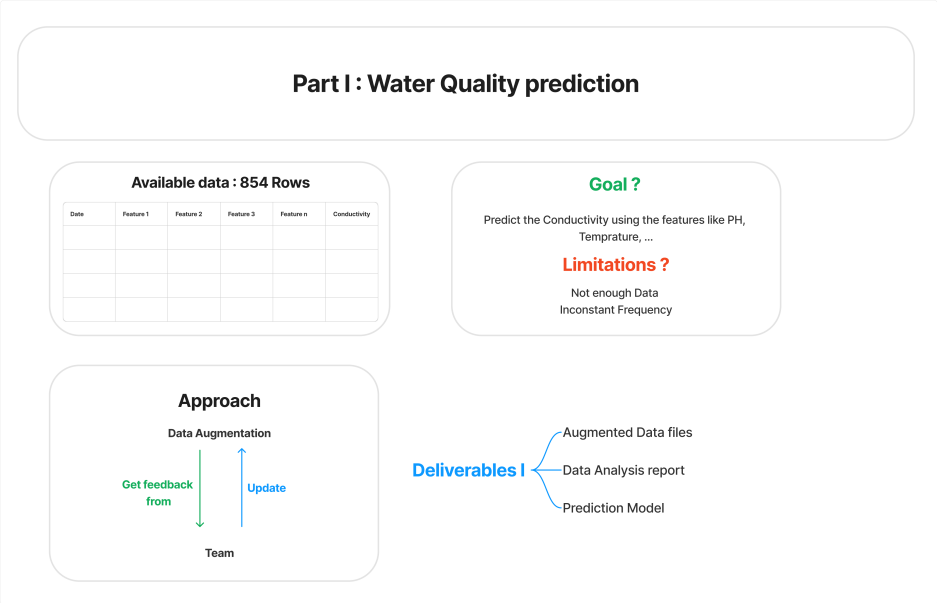

Figure 3.1: Workflow of the Water Quality Prediction task

3.2.1 Motivation and Objectives

RO systems are widely used to purify water by removing salts, contaminants, and other impurities, but their performance can be affected by various factors, including changes in feed water quality, membrane fouling, and operational conditions. By accurately predicting water quality, operators can ensure that the RO systems perform optimally, minimizing the risk of producing substandard water and reducing operational costs.

This research was initiated at the request of ONEE to better understand the factors that most influence water quality in RO systems. One of the key indicators of water quality in this context is the conductivity level, which measures the ability of water to conduct electricity—a direct proxy for the concentration of dissolved salts and impurities. For effective water purification, the conductivity level in the permeate (purified water) should be relatively low. High conductivity indicates that the RO membrane is allowing too many salts and other contaminants to pass through, suggesting a decline in performance.

The primary objectives of this chapter are:

1. **Develop Machine Learning Models for Water Quality Prediction:** To create models that can accurately predict key water quality parameters, especially conductivity levels,

in RO systems. This involves analyzing input data from sensors and operational logs to identify patterns and factors that significantly affect conductivity.

2. **Identify Influencing Factors:** To determine the factors that most influence water quality, particularly conductivity, in RO systems. Understanding these factors will help in optimizing operational parameters and improving RO system performance.

3. **Enhance Operational Efficiency and Reliability:** To use predictive models to anticipate changes in water quality and adjust operational parameters accordingly, thereby improving the overall efficiency and reliability of RO systems.

3.2.2 Data collection and pre-processing

Features

DATE_HEURES	Precarts_PI203	Sortie_pompe	Primaire_PI400	Secondaire_PI302	Tertiaire_PI403	perm_PI304	PermFIT304	Conc_FIT403	Entrée_AIT206	Perm_AIT304	conc_mesur	T°C_TIT100	Perm_PH_mesur	Entrée_PHAIT204
28/11/2012	4,3	11,73	10	9,5	8	0,5	47,1	16,3	961	20	NaN	18,8	NaN	7
29/11/2012	4,3	12,08	10	9,7	8,7	0,5	48,5	17,3	991	19	NaN	18,8	NaN	7
08/12/2012	4,8	12,06	9,7	9,5	8,78	0,5	47	15,6	934	17	NaN	16,8	NaN	6,8
13/12/2012	4,2	12,02	11,5	10,6	9,2	0,5	45,4	17,3	934	17	NaN	16,5	NaN	7,06
18/12/2012	4,4	12,6	11,3	10,5	9	0,5	47,6	15,9	958	17	NaN	16,8	NaN	7,1

Figure 3.2: RO Data from ONEE Kenitra.

Here is a brief explanation of the parameters in the dataset:

1. **DATE_HEURES:** This represents the date and time of the measurement, indicating when the data was collected.

2. **Precarts_PI203:** This parameter represents the pressure differential across a specific section of the RO system (such as a membrane or filter), measured in bars. It is used to monitor the performance and detect fouling or clogging.

3. **Sortie_pompe:** This is the output pressure of a pump in the RO system, indicating the pressure at which water is being fed into the system or through a particular section.

4. **Primaire_PI400:** This parameter represents the primary stage pressure within the RO system, showing the pressure level in the initial stage of water treatment.

5. **Secondaire_PI302:** This is the secondary stage pressure, indicating the pressure level in the second stage of the RO process.

6. **Tertiaire_PI403:** This parameter denotes the tertiary stage pressure, which is the pressure level in the third stage of the RO system.

7. **perm_PI304:** This refers to the permeate pressure, which is the pressure of the purified water (permeate) that passes through the RO membranes. Lower permeate pressure can indicate effective water purification.

8. **PermFIT304:** This represents the permeate flow rate, measured in cubic meters per hour (m^3/h). It shows the volume of water being purified and passing through the system.

9. **Conc_FIT403:** This parameter is the concentrate flow rate, indicating the volume of water that contains the rejected salts and impurities. It is measured in m^3/h and helps assess the system's efficiency.

10. **Entrée_AIT206:** This is the feed water conductivity, measuring the ability of the water entering the RO system to conduct electricity, which correlates with the concentration of dissolved salts and impurities.

11. **Perm_AIT304:** This represents the conductivity of the permeate water, which indicates the quality of the purified water. Lower conductivity means better water quality with fewer dissolved salts.

12. **conc_mesur:** This parameter indicates the concentration of salts and impurities in the concentrate or brine stream of the RO system, helping to monitor the rejection rate of contaminants.

13. **T°C_TIT100:** This represents the temperature of the water within the RO system, which can affect membrane performance and water quality.

14. **Perm_PH_mesur:** This parameter measures the pH level of the permeate water, indicating its acidity or alkalinity. It is crucial for assessing water quality and the potential impact on distribution systems.

15. **Entrée_PHAIT204:** This is the pH level of the feed water entering the RO system, which can influence the efficiency of the RO process and membrane lifespan.

For our research, the primary focus is on the conductivity parameter, specifically the conductivity of the permeate water (Perm_AIT304), as it is a key indicator of water quality in reverse osmosis (RO) systems. Conductivity will be the target variable in our predictive models. Our goal is to accurately forecast changes in water quality by predicting this parameter based on other operational data. Additionally, we will not consider time (DATE_HEURES) as a predictor variable in our models, as our focus is on understanding the relationships between the operational parameters and water quality, rather than on temporal trends.

Data cleaning and pre-processing

During the data cleaning step, we performed several operations to prepare the dataset for machine learning analysis.

First, we converted all relevant numerical data into float format to ensure consistent data types across the dataset, which is essential for accurate calculations and model training.

Next, we addressed the issue of missing values, which can negatively impact the performance of machine learning models. To handle this, we replaced all null values with the mean of their respective columns. The decision to replace missing values with the mean of their respective columns was made after consulting with experts in the field who emphasized that using the mean was a suitable approach because it preserves the overall distribution and typical behavior of the data, reflecting normal operational conditions. By filling in missing values with the mean, we ensured that the dataset remains representative of typical system performance, which is crucial for developing accurate and reliable predictive models. This expert input helped us make informed choices during the data cleaning process, aligning our methodology with industry best practices.

Finally, we dropped columns that were empty as they wouldn't be helpful in the prediction process.

After completing these cleaning steps, we saved the newly cleaned dataset in a new CSV file, ensuring that it is ready for subsequent analysis and model development.

Precarts_PI203	Sortie_pompe	Primaire_PI400	Secondaire_PI302	Tertiaire_PI403	perm_PI304	PermFIT304	Conc_FIT403	Entrée_AIT206	Perm_AIT304	T°C_TIT100	Entrée_PHAIT204
4.300000	11.73	10.0	9.5	8.00	0.500000	47.1	16.3	961.000000	20.0	18.8	7.00
4.300000	12.08	10.0	9.7	8.70	0.500000	48.5	17.3	991.000000	19.0	18.8	7.00
4.800000	12.06	9.7	9.5	8.78	0.500000	47.0	15.6	934.000000	17.0	16.8	6.80
4.200000	12.02	11.5	10.6	9.20	0.500000	45.4	17.3	934.000000	17.0	16.5	7.06
4.400000	12.60	11.3	10.5	9.00	0.500000	47.6	15.9	958.000000	17.0	16.8	7.10
...	...	...	...	...	...	...	...	...	...	...	...
4.200000	15.87	14.5	14.0	13.00	0.600000	41.2	16.5	1204.000000	15.0	17.4	7.90
4.500000	15.10	13.9	13.3	12.10	0.600000	40.4	16.5	1137.000000	17.0	18.9	7.90
3.600000	7.98	10.5	10.0	9.00	0.500000	44.5	20.9	1079.591805	21.4	18.9	5.90
3.600000	8.00	9.5	8.8	8.00	0.592306	44.4	33.6	98.000000	0.0	22.6	7.60
4.312833	7.79	8.9	8.4	7.00	0.592306	44.0	34.0	98.000000	0.0	23.4	7.70

Figure 3.3: Clean RO data from ONEE Kenitra.

Correlation

The figure illustrates the correlation matrix for the parameters in the dataset, using a color gradient to indicate the strength and direction of the relationships between them. Darker red colors suggest strong positive correlations, while darker blue colors indicate strong negative correlations, with lighter colors representing weaker or no correlations.

It shows a strong negative correlation with "conc_mesur" and "Entrée_AIT206," indicating that as the values of these parameters increase, the value of "Perm_AIT304" tends to decrease.

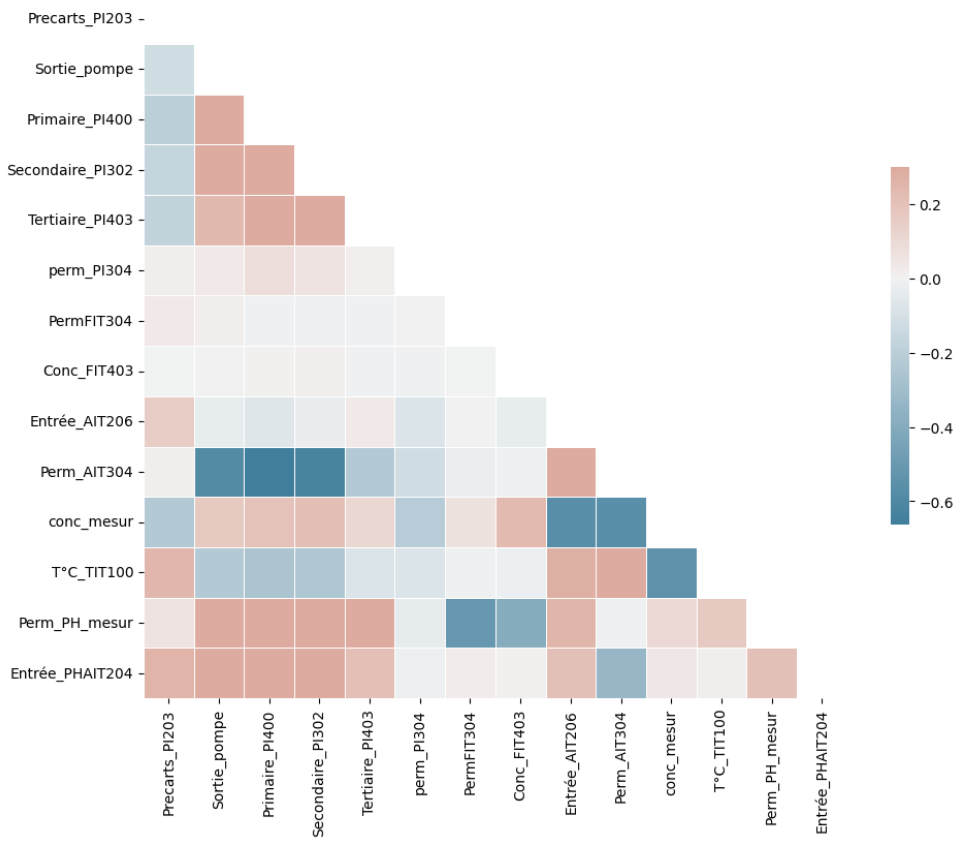

Figure 3.4: RO data from ONEE Kenitra correlation matrix.

Additionally, “Perm_AIT304” has a moderate negative correlation with “Sortie_pompe” and “Primaire_PI400,” suggesting that these parameters also inversely affect “Perm_AIT304” to a lesser extent. These correlations are important as they provide insights into how different factors within the reverse osmosis system interact with “Perm_AIT304,” which is crucial for predicting water quality outcomes and optimizing the system’s performance.

3.2.3 Results and Discussion

Random Forest algorithm

Random forest is a commonly used machine learning algorithm, trademarked by Leo Breiman and Adele Cutler, that combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems [30]. Random Forests are resilient to overfitting, especially when dealing with datasets with many variables, by averaging out the predictions of multiple uncorrelated trees.

This algorithm is also robust against noisy data, making it suitable for real-world applications where data might be incomplete or contain errors.

Algorithm used	Random Forest
Testing size	20%
Mean Squared Error	5.528
Mean Absolute Error	1.238

Table 3.1: Prediction Results of Random Forest on data from ONEE kenitra.

XGboost algorithm

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way [31]. XGBoost builds an ensemble of trees sequentially, where each tree tries to correct the errors of the previous one by focusing on the residuals. This approach allows XGBoost to perform exceptionally well in predicting conductivity by effectively capturing complex, non-linear relationships and interactions between various parameters like flow rates, pressures, and temperatures. Additionally, its ability to handle missing data and support parallelization makes it a fast and accurate option for predicting water quality metrics such as conductivity.

Algorithm used	XGBoost
Testing size	20%
Objective	reg:squarederror
Mean Squared Error	4.777
Mean Absolute Error	1.240

Table 3.2: Prediction Results of XGBoost on data from ONEE kenitra.

Linear Regression algorithm

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable [32]. While simpler

than Random Forest and XGBoost, Linear Regression can still be valuable in conductivity prediction, especially if the relationship between the features and conductivity is approximately linear. This algorithm is easy to interpret, making it a good choice for understanding the direct impact of each parameter on conductivity. However, its simplicity means it might not capture complex interactions or non-linear relationships as effectively as the other algorithms.

Algorithm used	**Linear Regression**
Testing size	20%
Mean Squared Error	2390.791
Mean Absolute Error	6.271

Table 3.3: Prediction Results of Linear regression on data from ONEE kenitra.

Discussion

According to Table 3.2, the results show that the XGBoost algorithm has the best performance for predicting conductivity in RO systems, with the lowest Mean Squared Error (MSE) of 4.777 and a Mean Absolute Error (MAE) of 1.240, showing its high accuracy in capturing the complex relationships in the data. In Table 3.1, the Random Forest algorithm also performs well, with an MSE of 5.528 and an MAE of 1.238, suggesting it effectively handles non-linear patterns, although slightly less accurately than XGBoost. In contrast, as shown in Table 3.3, the Linear Regression model shows significantly worse performance, with a very high MSE of 2390.791 and an MAE of 6.271, reflecting its inability to model the non-linear relationships in the dataset effectively. These results highlight that for predicting water quality, specifically conductivity, in RO systems, tree-based algorithms like XGBoost and Random Forest are much more effective than Linear Regression, which fails to account for the complexity of the data.

3.3 Predictive maintenance using machine learning

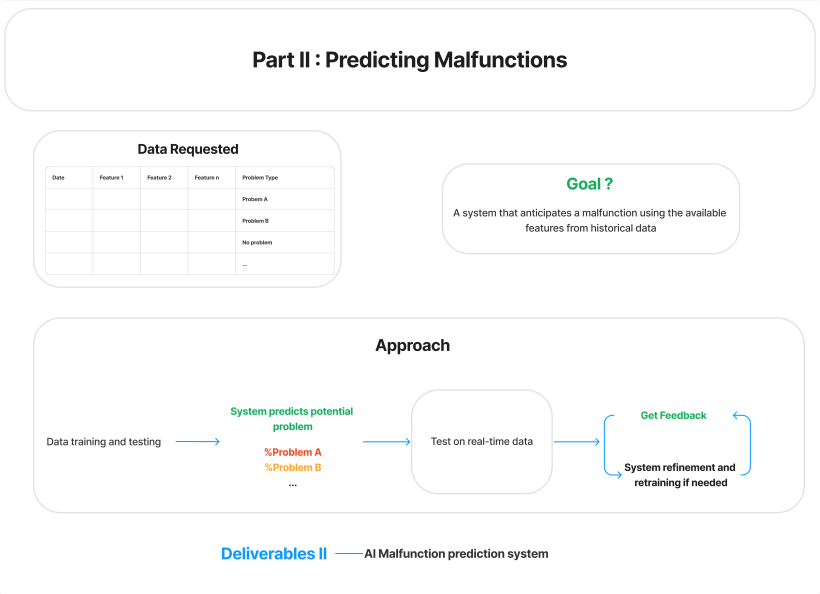

Figure 3.5: Workflow of the task : Predicting Malfunctions within RO systems

3.3.1 Motivation and Objectives

Predictive maintenance using machine learning is a crucial component of modern industrial operations, especially in systems like reverse osmosis (RO) that are sensitive to operational inefficiencies and downtime. The primary motivation behind implementing predictive maintenance in RO systems is to optimize the lifespan and performance of the equipment, minimize unexpected breakdowns, and reduce maintenance costs. Traditional maintenance approaches, such as reactive or scheduled maintenance, often lead to unnecessary downtime and can result in significant costs due to either premature maintenance actions or unexpected equipment failures. By leveraging machine learning algorithms, we can analyze historical data and detect patterns that indicate potential failures.

The objectives of implementing predictive maintenance using machine learning in RO systems are threefold: firstly, to enhance the reliability and efficiency of the systems by predicting failures or performance degradation; secondly, to reduce operational costs associated with unplanned downtimes and repairs by going from reactive to proactive maintenance strategies; and thirdly, to improve decision-making processes by providing data-driven insights into the operational status and maintenance needs of the equipment. These objectives aim to ensure the continuous and optimal operation of RO systems, thereby contributing to higher productivity and better resource management.

3.3.2 Data collection and pre-processing

For this section of the project, we obtained data from ONEE Khenifra, which provided us with operational records from their reverse osmosis (RO) systems. The data contained multiple types of issues related to system performance, however, to streamline the predictive maintenance model, it was necessary to simplify these various problem types into a more manageable format for binary classification. We manually labeled the data, categorizing each entry into one of two states: “problem” (represented as 1) or “no problem” (represented as 0). This binary approach allowed us to focus on the fundamental task of identifying whether a problem exists, rather than predicting specific problem types, thus making the model more effective and efficient in a real-world predictive maintenance context. The pre-processing phase also included cleaning the data, handling missing values, and normalizing the feature set to ensure consistency and accuracy in the model’s predictions.

Correlation

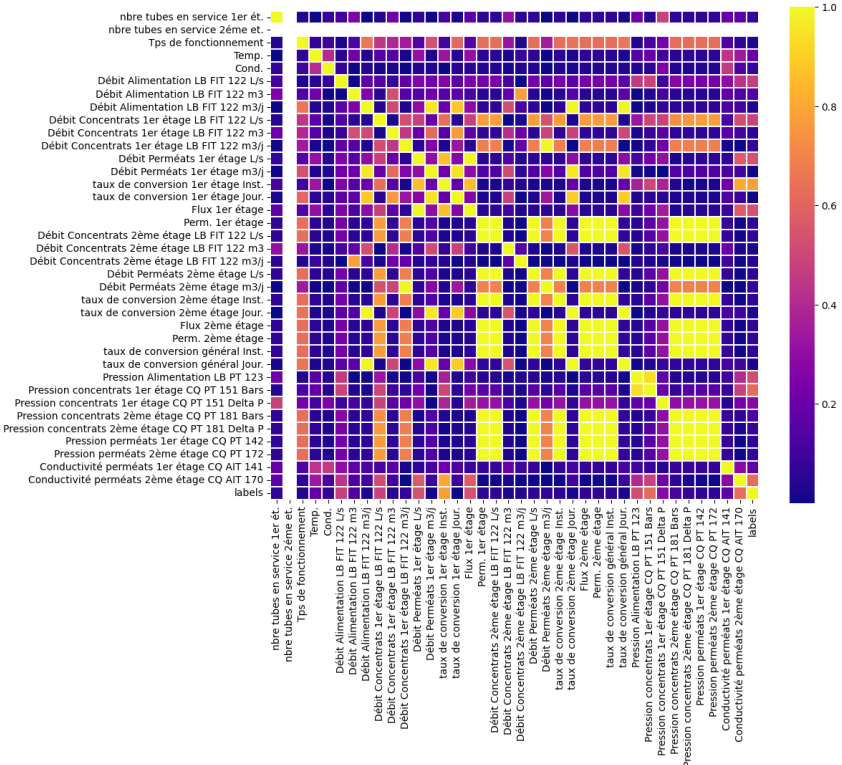

Figure 3.6: RO data from ONEE Khenifra correlation matrix.

The figure shows the correlation between various parameters of the reverse osmosis (RO) system and the binary labels indicating the presence or absence of a problem. From the heatmap, it is evident that parameters such as “Conductivité perméats 1er étage CQ AIT 141” (Permeate Conductivity of Stage 1), “Conductivité perméats 2ème étage CQ AIT 170” (Permeate Conductivity of Stage 2), “Temp.” (Temperature), and “Débit Concentrats” (Concentrate Flow Rates) have a significant impact on the labels. These parameters show a higher correlation with the label, suggesting they are critical indicators for predicting the presence of issues in the RO system. This information is valuable for focusing predictive maintenance efforts on the most influential factors affecting system performance.

3.3.3 Results and Discussion

The Random Forest and XGBoost models were applied to predict problems in the reverse osmosis (RO) system using a testing size of 20%, and both showed promising results in terms of predictive accuracy. In Table 3.4, The Random Forest model achieved a Mean Squared Error (MSE) of 0.036 and a Mean Absolute Error (MAE) of 0.055, indicating a reasonable level of accuracy, though the slight difference between MSE and MAE suggests occasional larger deviations. On the other hand, according to the results in Table 3.4, the XGBoost model also achieved an MSE of 0.036 but with a lower MAE of 0.036, highlighting its consistency and ability to make more precise predictions without significant outliers. The XGBoost model was optimized with specific hyperparameters, such as a subsample rate of 0.8, 500 estimators, a minimum child weight of 2, a maximum depth of 20, a learning rate of 0.4, gamma of 0.001, and a column sample by tree of 0.6, which contributed to its superior performance. Furthermore, XGBoost demonstrated high precision and recall for both classes, with precision and recall values of 0.97 and 0.98 for class 0.0 (no problem) and 0.96 and 0.94 for class 1.0 (problem), resulting in F1-scores of 0.97 and 0.95, respectively. This balanced performance across both classes suggests that XGBoost is slightly more robust than Random Forest for the predictive maintenance of RO systems, offering better generalization and reliability in identifying potential issues.

Algorithm used	Random Forest
Testing size	20%
Mean Squared Error	0.036
Mean Absolute Error	0.055

Table 3.4: Prediction Results of Random Forest on data from ONEE Khenifra.

Algorithm used	XGBoost
Testing size	20%
Objective	binary:logistic
Mean Squared Error	0.036
Mean Absolute Error	0.036

Table 3.5: Prediction Results of XGBoost on data from ONEE Khenifra.

3.4 Limitations

One of the major limitations encountered in this study was the insufficient amount of data available for implementing sequential models, such as Long Short-Term Memory (LSTM) networks. Sequential models like LSTM are designed to analyze and predict patterns in time series data, where the sequence and timing of data points play a crucial role. These models are effective in scenarios where past states significantly influence future outcomes, making them ideal for predicting system behaviors over time. However, due to the limited quantity and temporal resolution of the data provided for both parts of our study, we were unable to establish the necessary sequences that would allow LSTM or similar algorithms to operate effectively. This restriction prevented us from fully leveraging the potential of sequential models to capture long-term dependencies and trends within the data, which could have enhanced the accuracy and robustness of our predictive maintenance efforts.

Chapter 4: Troubleshooting of problems related to RO systems using LLMs

4.1 Introduction

This chapter focuses on the use of Large Language Models (LLMs) for troubleshooting problems related to Reverse Osmosis (RO) systems. RO systems are widely used in water treatment facilities for their ability to remove impurities and provide clean water. However, maintaining the optimal performance of these systems can be challenging due to various operational issues, such as membrane fouling, fluctuations in feedwater quality, and equipment malfunctions. Traditional troubleshooting methods often rely on extensive human expertise and can be time-consuming, especially when dealing with complex or uncommon issues. By applying LLMs, we aim to streamline the troubleshooting process, providing rapid and accurate solutions based on comprehensive data analysis and knowledge extraction from multiple reliable sources, including technical manuals and historical data. This approach not only improves response times but also enhances the consistency and quality of the troubleshooting process.

4.2 Motivation and Objectives

4.2.1 Motivation

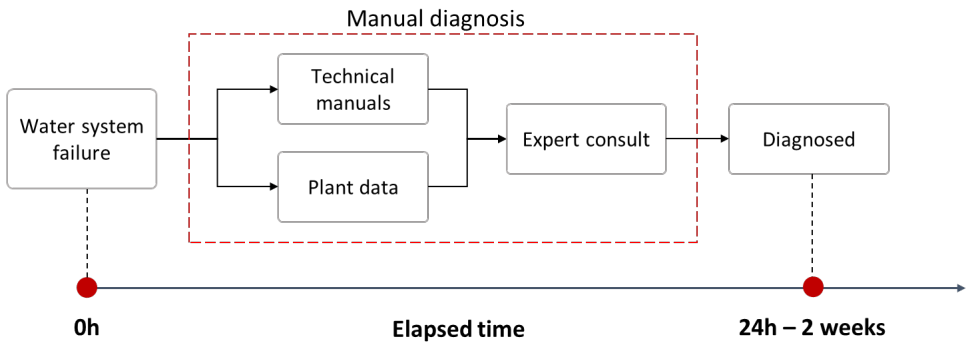

Figure 4.1: Illustration of how much time it takes to diagnose an issue occurring in a RO system

The motivation for employing LLMs in troubleshooting RO systems comes from the high complexity of these systems and the demand for more efficient maintenance strategies. RO systems are essential to many industrial and water treatment applications, where downtime performance can have significant consequences as shown in the figure below. Traditional troubleshooting often involves sifting through vast amounts of technical documentation and data logs, which can be inefficient and prone to human error. Additionally, the availability of experienced technicians who can diagnose and resolve these issues quickly is often limited, especially in remote or under-resourced locations. The ability of LLMs to understand and process natural language enables them to identify and diagnose issues effectively, providing actionable recommendations that can be implemented swiftly, thus minimizing downtime and ensuring the continuous operation of RO systems.

4.2.2 Objectives

The primary objective of this part is to develop an advanced troubleshooting framework for RO systems using LLMs. To achieve this, several key goals have been established as shown in Figure 4.2.

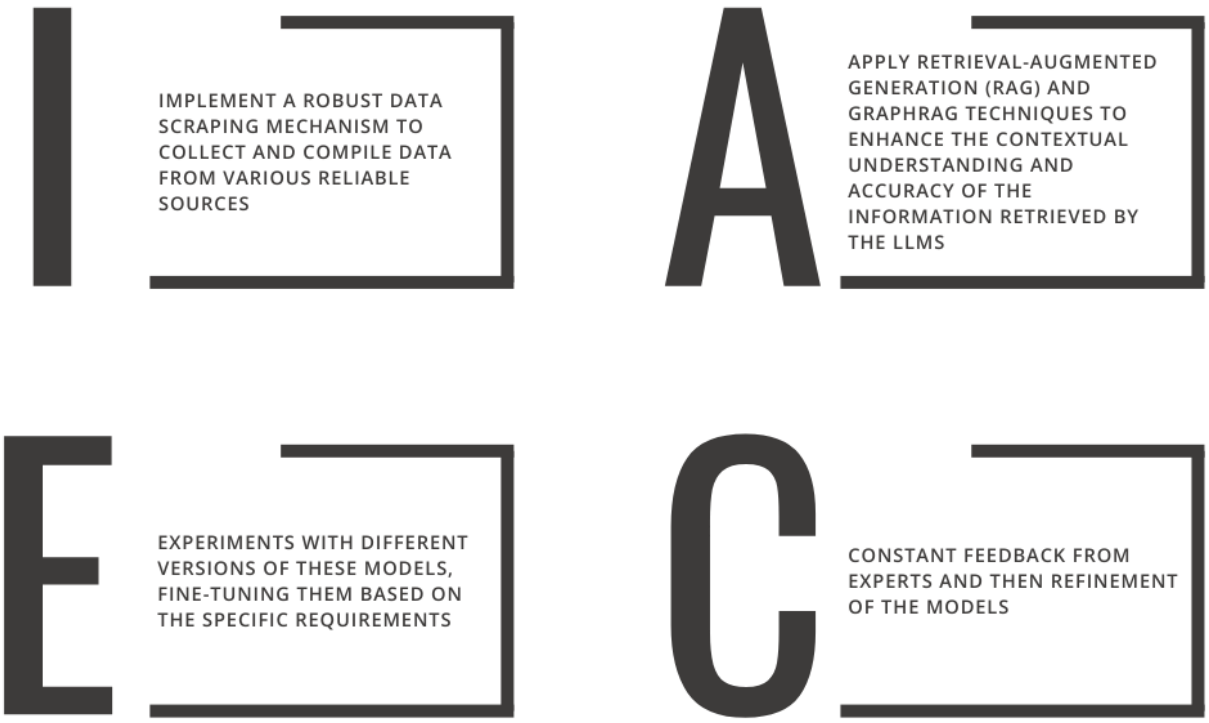

Figure 4.2: Objectives of using LLMs to troubleshoot RO-related issues

- Implement a robust data scraping mechanism to collect and compile data from various reliable sources
- Apply Retrieval-Augmented Generation (RAG) and GraphRAG techniques to enhance the contextual understanding and accuracy of the information retrieved by the LLMs
- Experiments with different versions of these models, fine-tuning them based on the specific requirements
- Constant feedback from experts and then refinement of the models

First, we aim to implement a robust data scraping mechanism to collect and compile data from various sources, including technical manuals, operational logs, and other relevant documentation suggested by the Aquadviser team. This step is crucial for creating a comprehensive dataset that the LLMs can use to learn and infer from.

Second, we need to apply Retrieval-Augmented Generation (RAG) and GraphRAG techniques to improve the contextual understanding and accuracy of the information retrieved by the LLMs. These methods combine the retrieval of relevant data with the generation of responses, enabling the models to provide more accurate and contextually relevant troubleshooting advice. Third, we intend to run a series of experiments with different versions of these models, fine-tuning them

based on the specific requirements of RO systems and evaluating their performance through expert feedback. This iterative process will involve assessing the models' ability to identify issues, suggest appropriate corrective actions, and provide clear explanations that are easily understood by operators and technicians.

Ultimately, the goal is to create a reliable, AI-driven troubleshooting system that can enhance the efficiency and effectiveness of RO system maintenance, reduce downtime, and improve water treatment outcomes.

4.3 Data collection and pre-processing methods

4.3.1 Tools : Selenium for data scrapping

Selenium is a powerful tool for controlling web browsers through programs and performing browser automation. It is functional for all browsers, works on all major OS, and its scripts are written in various languages i.e Python , Java , C# , etc, we will be working with Python [33].

Selenium was chosen for this project due to its robust capabilities for automating web browsing and data scraping. Given the diverse range of online sources from which we needed to extract information, Selenium provided the flexibility and control necessary for navigating complex websites, interacting with dynamic content, and efficiently capturing the required data.

One of the main reasons for using Selenium is its ability to handle websites that require user interaction, such as logging in, filling out forms, or clicking through multiple pages to access specific data. These tasks are challenging for simpler scraping tools that are not designed to handle JavaScript-heavy or interactive web pages. Selenium, however, operates by automating browser actions in real-time, allowing it to mimic human interaction closely. This feature was particularly useful for collecting data from technical forums, online manuals, and documentation pages that required user authentication or contained dynamic content that could not be easily accessed through static scraping methods.

Overall, Selenium was instrumental in our ability to gather comprehensive and accurate datasets from a wide range of online sources, forming a solid foundation for the subsequent application of machine learning models and Large Language Models (LLMs) in troubleshooting RO system problems.

4.3.2 Workflow

This general workflow for data scrapping using Selenium involves several key steps to ensure efficient and automated extraction of information from web-based sources. First, the environment

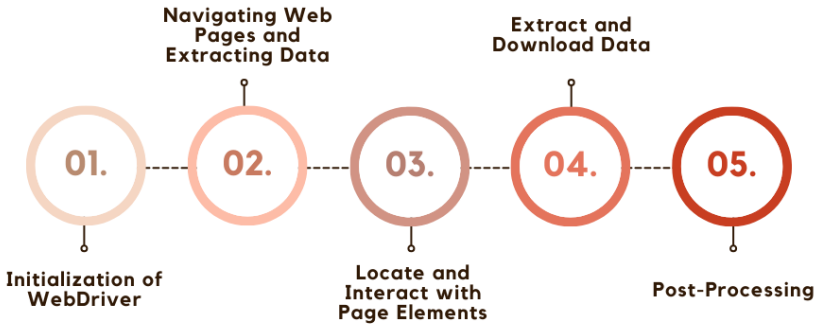

Figure 4.3: Workflow of data scrapping using Selenium

is set up by installing Selenium and configuring the necessary WebDriver, such as ChromeDriver, to control the browser. The WebDriver is then initialized to start a new browser session, and it navigates to the target webpage containing the desired data. Once on the webpage, Selenium's explicit wait functions are employed to ensure that all elements are fully loaded and ready for interaction. The workflow then focuses on locating specific elements, such as headers, buttons, or tables, that need to be interacted with to reveal or access the data. Using commands to scroll, click, or expand sections, Selenium makes these elements visible or initiates downloads of necessary files. During this process, error-handling mechanisms are crucial to manage any potential issues, such as missing elements or failed downloads, ensuring a robust data collection procedure. After all relevant data has been collected and downloads are complete, the browser session is closed to release resources. Finally, post-processing is performed on the collected data, such as organizing files and preparing them for further analysis.

This comprehensive workflow allows for automated, repeatable, and reliable data extraction from dynamic web pages.

4.4 Rag integration

4.4.1 Introduction

In this section, we describe the comprehensive experiments conducted over a two-month period, during which we explored different language models and fine-tuned their parameters to optimize performance for specific tasks related to water desalination systems. The data used in these experiments were collected by scraping various desalination-related websites recommended by Aquadviser. This dataset includes a diverse range of materials such as case studies, data sheets, Q&A files, and technical manuals, all of which are pertinent to the field of reverse osmosis (RO) systems.

Our approach began with the implementation of the open-source LLM, Llama3, with Retrieval-Augmented Generation (RAG), which we applied first to the entire collection of documents and then exclusively to the Q&A file provided by Dupont. We also experimented with the newly released Gemini model on the Q&A file to compare its performance. Following these initial trials, we decided to shift our focus to a more robust language model, OpenAI's GPT-4o. To enhance the relevance and accuracy of the responses, we employed Retrieval-Augmented Generation (RAG) on a single, most critical document as declared by Aquadviser, as it contains most of the answers: Filmtech: the technical manual.

The effectiveness of these models and configurations was assessed by applying them to a set of 20 diverse RO-related questions that were carefully crafted by the Aquadviser team to cover a broad spectrum of scenarios and challenges encountered in desalination processes. Through iterative feedback and continuous engineering of new solutions based on these evaluations, we refined our approach to better fit the needs of the industry, ultimately leading to a deeper understanding of how these advanced language models can be leveraged for technical and operational improvements in desalination systems.

4.4.2 Using Llama3

On Q&A file Dupont

Model name	Llama-3-8B-it
Open source	Yes
Embedding	google/text-embedding-004
Data	Q&A Dupont
Top_k	1
System_Prompt	As an assistant for question-answering tasks specializing in Reverse Osmosis (RO) Water Treatment Systems. Your goal is to provide accurate, concise, and helpful information, while including URLs mentioned in the Context as references within the text. Ensure your answers are direct, clear, logical, and tailored to the user's question, aiming to solve their problem effectively. If additional information is needed, indicate this explicitly
LLM Temperature	0.9

Table 4.1: Configuration details for the Llama-3-8B-it model used on the Q&A file Dupont.

The table above describes the first model we experimented with, that is Llama-3-8B-it. Using the open-source Llama-3 variant allows for customization, while the google/text-embedding-004 model provides high-quality embedding that enhance contextual understanding. The data sourced from Dupont's Q&A file ensures domain-specific knowledge, and setting Top_k: 1 focuses on delivering the most relevant answer. The prompt template directs the model to provide accurate and concise information with necessary references, and a temperature setting of 0.9 strikes a balance between generating diverse responses and maintaining precision.

Feedback and discussion

According to feedback made by Aquadvisers, the model only got 5 answers out of 20 that were fully correct. Several key issues were noted in the model's performance according to the team:

1. **Accuracy and Relevance:** Many of the answers provided by the model were either wrong or lacked relevance to the questions posed. For instance, the model often missed important information that was clearly available in the source documents. It sometimes provided details that were not requested or deviated from the specific context of the question, indicating a need for better contextual understanding and focus.
2. **Linking and Source Use:** There were several instances where the model provided incorrect links or referenced irrelevant materials. This shows a significant gap in how the model utilizes source documents to extract and link appropriate information. For example, it sometimes linked to sections not related to the query topic, such as providing a link on sanitization when the question was about something else entirely.
3. **Detail and Precision:** The feedback also pointed out that even when the model provided correct answers, it often lacked detail and failed to present concise, clear, and specific information. In many cases, the model could have benefited from extracting more detailed data from the documents to enrich its responses.
4. **Understanding of Context:** In several answers, the model misunderstood the context or failed to address the question's core intent. For example, it answered questions about membrane specifications without focusing on the relevant parameters, or it did not ask clarifying questions when the original query was vague.

On all documents

Model name	llama-3-70b-instruct
Open source	Yes
Embedding	BAAI/bge-small-en-v1.5
Data	All scrapped data including the Q&A file
Top_k	5
System_Prompt	You are an expert in reverse osmosis with extensive knowledge from technical manuals. Provide precise and concise answers to the user's questions without mentioning the sources or documents. Make sure to answer precise questions related to calculations related to reverse osmosis.
LLM Temperature	0.01

Table 4.2: Configuration details for the Llama-3-70b-instruct model used on all documents.

The table above outlines the configuration for the llama-3-70b-instruct model, which was used to handle data from all scraped sources, including the Q&A file. The model uses the BAAI/bge-small-en-v1.5 embedding, which is designed to capture the semantic meanings of text effectively, ensuring that the model understands and utilizes the content accurately. The setting of Top_k: 5 suggests that the model generates multiple answers, from which the best 5 answers are selected, providing a range of possible responses to enhance accuracy. The prompt template is tailored for precise, concise responses, emphasizing expertise in reverse osmosis and requiring the model to provide clear answers without referencing source materials. The LLM temperature is set to 0.01, indicating a high level of determinism in the model's outputs, focusing on generating highly accurate and consistent responses with minimal variation, which is ideal for technical and calculation-related questions.

Feedback and discussion

According to feedback from Aquadviser, We noticed that the model only got 7 answers out of 20 that were fully correct. The feedback on this model highlights several key issues with the accuracy and relevance of the answers generated from the Q&A file related to reverse osmosis systems. The model was evaluated on its ability to provide correct and precise answers, with significant emphasis on using the right documents and delivering information that directly addressed the queries. It is evident from the comments that while the model occasionally produced correct answers, it frequently provided incorrect or imprecise responses. In many cases, the model failed to reference essential information clearly available in the technical manuals and FAQs from DuPont, indicating a gap in its ability to utilize the most relevant sources effectively. For example, in several instances, the model either missed key data points or mixed information from different applications, leading to inaccurate or irrelevant answers. There were also comments about the model providing answers that, while correct, lacked specificity or sufficient detail, suggesting that the model needs to better contextualize its responses based on the exact requirements of the questions.

Next steps

The feedback on both models suggests that while some answers were technically correct, they could have been improved by focusing more on the required information and eliminating unnecessary details. It also indicates that enhancing the model's ability to extract and synthesize relevant information from the documents accurately is crucial for future iterations. Additionally, Aquadviser team suggested to focus on one big document, which usually contains most answers to the 20 questions.

Our next steps involved introducing the state-of-the-art models GPT-3 and GPT-4o to enhance the quality and accuracy of our answers. Both models, known for their superior language understanding and contextual awareness, were chosen to address the shortcomings identified in the initial model, particularly in providing more relevant and precise responses. Alongside this, we also implemented improved embedding techniques to better capture the semantic relationships within the data. These embeddings were designed to enhance the model's ability to understand and utilize the content from documents more effectively. By upgrading to GPT-4o and refining our embedding strategies, we aimed to significantly boost the model's performance, ensuring that it generates answers that are both accurate and contextually appropriate, thereby meeting the specific needs of the reverse osmosis water treatment domain.

4.4.3 OpenAI GPT3.5-turbo on technical manual

In this part, we transitioned to using OpenAI's GPT3.5-turbo model via Azure, which provided us with API access to leverage this state-of-the-art language model. This switch was motivated by the need to enhance the accuracy and relevance of the model's responses to technical queries regarding reverse osmosis systems. We conducted a series of experiments (1, 2, and 3) with GPT3.5-turbo, each designed to implement specific improvements and optimizations. These experiments aimed to refine the model's understanding of the technical manual: 2021 -DuPont FilmTec™ RO Technical Manual and improve its ability to generate precise, contextually appropriate answers. The details of each experiment and the corresponding enhancements will be elaborated on in this section.

Experiment 1: Using GPT3.5-turbo [Model 1]

Model name	gpt-35-turbo-instruct
Open source	No
Embedding	text-embedding-ada-002
Data	2021 - DuPont FilmTec™ RO Technical Manual
Top_k	10
PDF Parser	pypdf
System_Prompt	As an assistant for question-answering tasks specializing in Reverse Osmosis (RO) Water Treatment Systems. Your goal is to provide accurate, concise, and helpful information. Ensure your answers are direct, clear, logical, and tailored to the user's question, aiming to solve their problem effectively. If additional information is needed, indicate this explicitly. Use the given context to answer the question.
LLM Temperature	0.7

Table 4.3: Configuration details for the gpt-35-turbo-instruct model used on the 2021 DuPont FilmTec™ RO Technical Manual.

The table above outlines the configuration parameters for using the gpt-35-turbo-instruct model, a large language model known for its advanced capabilities in understanding and generating human-like text. This model leverages text-embedding-ada-002 for embedding, which is a

powerful tool for capturing the semantic relationships within the text, enhancing the model's ability to comprehend and utilize technical content effectively. The data used in this configuration is the 2021 DuPont FilmTec™ RO Technical Manual, ensuring that the model is grounded in domain-specific knowledge. The Top_k parameter is set to 10, which means the model generates the top 10 most relevant responses, allowing for a broader exploration of potential answers and enhancing the likelihood of retrieving highly accurate information. For parsing the PDF of the technical manual, pypdf is employed, a Python library that facilitates efficient extraction and handling of text from PDF files, ensuring that the content is accurately fed into the model for analysis. The prompt template is tailored to establish the model as an assistant specializing in RO water treatment systems, focusing on providing clear, concise, and logical answers directly addressing the user's queries. This template guides the model to use the given context effectively and indicates when additional information might be needed, aiming to deliver comprehensive and helpful responses. Lastly, the LLM temperature is set to 0.7, which balances creativity and precision in the model's outputs. A temperature of 0.7 allows the model to generate varied yet relevant responses, maintaining a degree of randomness to cover diverse possibilities while still focusing on accuracy and clarity in technical contexts.

Feedback and discussion

The feedback on the second model's performance, which involved transitioning to OpenAI's gpt-35-turbo-instruct highlights several key points about its strengths and areas for improvement. The evaluation focused on whether the answers provided by the model were "precise" or "loose" and whether they correctly referenced the technical manual. The comments indicate that while the RAG model often provided more precise and contextually relevant answers compared to the standalone LLM, there were still instances where the model failed to extract the correct information from the manual or misunderstood the questions. For example, in some cases, the answers were marked as wrong despite the correct approach, often due to referencing the wrong section of the document or omitting critical details like specific tables or figures. Furthermore, there were multiple instances where both models (LLM and LLM + RAG) failed to mention necessary context or provide specific answers that align with the technical manual, indicating a need for better document parsing and information retrieval.

Next steps

To address these issues and improve the model's performance, the next steps will involve integrating more advanced tools for data processing and retrieval. Specifically, we plan to implement Apache Tika as a PDF parser to enhance our ability to extract accurate and comprehensive information from complex documents like technical manuals. Tika's robust capabilities in text extraction will ensure that all relevant data, including embedded tables and figures, are correctly parsed and available for the model to reference. Additionally, we will incorporate the Semantic

Chunker as our text splitter to better segment the documents into meaningful chunks. This method will improve the retrieval process by ensuring that the model receives well-organized, contextually coherent chunks of text, leading to more accurate and relevant answers.

Experiment 2: Using GPT3.5-turbo With tika & Langchain's Semantic Chunker [Model 2]

In this experiment, we used the same parameters in Table3, with only two modifications:

- **Pdf parser:** Tika
- **Text splitter:** Langchain's Semantic Chunker

Feedback and discussion

> The feedback on this model reveals that while there have been some improvements in the precision of answers compared to previous models, there are still significant issues that need to be addressed. The evaluation indicates that the model was more effective in generating answers that are closer to what was required, with several responses categorized as "very good" or "good." This suggests that the modifications made have positively impacted the model's ability to provide relevant answers. However, there are also numerous cases where the model failed to extract or correctly interpret information from the technical manual, often missing crucial details or referencing incorrect sections. For instance, many of the correct answers were noted to be lacking in specificity or depth, while incorrect answers frequently resulted from the model either not finding or not using the relevant chunks from the documents.

One key observation is that the model's chunking mechanism needs further refinement. The comments point out several instances where valuable chunks of information were missed, indicating that the model's ability to segment and utilize document content is still not optimal. This directly impacts the accuracy of the answers, as seen in scenarios where the model should have referred to specific pages or sections of the technical manuals but failed to do so. Additionally, there were cases where the model generated correct answers but did not capture all necessary keywords or context, resulting in responses that, while technically accurate, were not fully aligned with the question's intent.

Next steps

To address the identified issues with this model, we decided to integrate AzureAIDocumentIntelligenceLoader as our PDF parser. This method is designed to provide high-resolution optical character recognition (OCR) capabilities, ensuring that even the most detailed and complex

elements within the technical manual, such as embedded tables, are accurately extracted and formatted.

In addition to this, we have decided to implement the MarkdownHeaderTextSplitter from LangChain as our text splitter. This tool will segment the documents based on markdown headers, allowing us to break down the content into logically organized chunks. By splitting the text in this manner, we aim to improve the model's ability to navigate through the documents and identify relevant sections more effectively. This structured approach will facilitate better contextual understanding and more precise retrieval of information, ultimately leading to more accurate and detailed answers in response to technical questions.

Experiment 3: With AzureAIDocumentIntelligenceLoader & MarkdownHeaderTextSplitter [Model 3]

In this experiment, we used the same parameters in the previous experiment, with only two modifications:

- **Pdf parser:** AzureAIDocumentIntelligenceLoader
- **Text splitter:** Langchain's MarkdownHeaderTextSplitter

Feedback and discussion

Overall 6/10. The answers are concise, however sometimes the following issues pop up:

- mention tables that are not included in the answer (Answers 6 and 9).
- No need to add at the beginning of the answer [As an assistant for question-answering tasks specializing in Reverse Osmosis (RO) Water Treatment Systems, let me provide you with..]
- no need to add this sentence at the beginning of the answer "As an assistant specializing in RO water treatment systems, I can provide some information on" (Answer 8). Sometimes confusing additional info (Answer 10).
- No need to mention this in the answer [As an AI, I do not have personal opinions.] (Answer 16).

Next steps

In the next steps, we decided to start implementing OpenAI GPT-4o on our technical manual and see if the performance of our model will be any better.

4.4.4 OpenAI gpt-4o on technical manual

Experiment 1: [Model 3] with gpt-4o

In this experiment, we used the same parameters in [Model 3], with only one modification, which is the utilization of gpt-4o instead of gpt3.

Feedback and discussion

Overall **8.5/10**, this is so far the model that provides good answers with valuable details and well written and honest true answers without additional made up insights. Clear concise and to the point and very good way of presenting the answers like in answer 6 with bullet points. Very coherent answers and very well logical and meaningful connection between insights.

Cost

In this experiment, we wanted to see how much will it cost to use gpt-4o with our parameters. For this reason, we decided to study the average of input tokens, average output tokens, average total tokens, and at last the average cost per query in US dollars.

- **Average of input tokens** : 4679.45
- **Average of output tokens** : 240.8
- **Average of total tokens** : 4900.95
- **Average of cost per query in Us dollars** : 0.026 $

Next steps

After several meetings with the team, we identified a critical improvement to address the model's limitations in accurately retrieving and referencing information: **splitting the entire document by chapter**. This approach was proposed because most of the answers required by the model are typically found within specific chapters of the technical manuals. By dividing the document into chapters, we aim to enhance the model's ability to focus on the most relevant sections, reducing the chances of irrelevant or incorrect information being included in the answers. While we recognize that this method will be **computationally expensive**, as it involves processing and analyzing each chapter individually, we believe that the potential benefits justify the investment. This approach is expected to significantly improve the precision and contextual accuracy of the model's responses, as it will be able to more effectively match the queries with the corresponding

chapters. The experiment with this chapter-based splitting will be important in determining whether this strategy can provide the level of detail and specificity needed for technical question-answering tasks.

Experiment 2: [Model 3] with gpt-4o and Chapter splitting approach

In this experiment, we used the same parameters in [Model 3], with two major modifications, which are :

- The utilization of gpt-4o instead of gpt3
- **The splitting method:** splitting the entire document by chapter as explained before.

Feedback and discussion

The answers are in general good and contain good details. If the answer of the question is in the manual, it finds the corresponding chunks (containing the answer) and rate it as the most similar chunk, hence it uses the corresponding entire chapter as the context. However, in some cases it added additional information which is not correct but also not mandatory to satisfy the answer. In cases where the answers are not in the manual technique, this model, sometimes, does not recognize and give a similar answer as the LLM (without RAG), in this case however, we expected to not give an answer and say it has no correct answer. In two cases (answer 11 and 12) where the answers were given wrong. The increase in answer size by containing the non necessary information/details may be due to updating based LLM from 03 to 4o. This update is however mandatory because of the size of the context, since 03 had a smaller limit of the maximal prompt size. (The answers are generally good, insightful and detailed. When a question's answer is in the DuPont membrane technical manual, the model identifies the relevant sections (containing the answer) and rates it as the most similar part, using the entire chapter as context. However, in some instances, it added extra information that is incorrect and unnecessary for answering the question. When the answers are not in the DuPont membrane technical manual, the model sometimes fails to recognize this and provides a similar answer as a standard language model (without Retrieval-Augmented Generation). In these cases, we expected it to indicate that there is no correct answer rather than provide one. In two cases (answers 11 and 12), the answers were wrong.

Next steps

The model generally did well, but we came to a conclusion that we can now experiment more with the prompt, and see how that will eventually affect our model. The prompt we will be using

should handle the issues related to how long the answers are, as well as being a bit more straight to the point.

Cost

In this experiment, we eventually had to use an entire chapter as context to out model, and so some concern was raised around the cost of this operation. For this reason, we decided to study the average of input tokens, average output tokens, average total tokens, and at last the average cost per query in Us dollars.

- **Average of input tokens** : 19911.0
- **Average of output tokens** : 240.8
- **Average of total tokens** : 20151.8
- **Average of cost per query in US dollars** : 0.103 $

Experiment 3: Apply a new prompt on the latest two models

In this final experiment with RAG and OpenAI gpt-4o, we will use an enhanced prompt for a better generation :

—Role—
As an assistant for question-answering tasks specializing in Reverse Osmosis (RO) Water Treatment Systems.
Use the following pieces of retrieved context to answer the question.
—Goal—
Your goal is to provide accurate, concise response.
If you don't know the answer, just say so. Do not make anything up, or add information that is not answering directly the user's question.
Ensure your answers are direct, clear, logical, and tailored to the user's question, aiming to solve their problem effectively.
Do not include information where the supporting evidence for it is not provided.
If additional information is needed, indicate this explicitly.
—Target response length and format—
single straight to the point paragraph, add a section called: Additional information, that should be longer, in which you give more details that the user might need.
Style the response in markdown.
—Context—
context

Feedback and discussion

> For the splitting by sections with enhanced prompting, it got an Overall 8.3/10, this is so far also the model that provides good answers with valuable details and well written and honest true answers without additional made up insights. Very coherent answers and very well logical and meaningful connection between insights. Slightly lack the presentation of bullet points and bold written points but it is so far one of the best. The splitting by chapter was also showcased as a great performance.

Since both models with the new enhanced prompt performed well, we decided to go with the one that objectively costs less, and that is the model with the chunking by sections. The choice was made after several meetings and discussions, and it was eventually the most optimal model to be deployed in our platform.

4.4.5 Discussion

Throughout this project, we conducted a series of experiments and refinements using different models to improve the accuracy and relevance of answers provided for technical questions related to reverse osmosis systems. The process began with the implementation of Llama3, which was tested both on the Q&A file from Dupont and across all relevant documents. Despite some successes, the feedback highlighted the limitations of this model, particularly in its ability to provide precise answers when drawing from multiple sources. The results showed that while Llama3 could handle certain queries effectively, it often struggled with the complexity and specificity required in a technical domain like water treatment. These limitations led to the decision to explore more powerful models.

We then moved on to experiments with OpenAI's GPT-3 on the technical manual. Over three separate experiments, we adjusted the model's parameters and retrieval methods, gradually improving its performance. The feedback indicated that GPT-3 was more capable of handling complex queries and retrieving relevant information. However, the model still occasionally provided responses that lacked the necessary depth or specificity, particularly when the answers were dispersed across different sections of the manual. Each iteration brought us closer to our goal, with improvements in how the model parsed and understood the technical content, but the results were still not fully satisfactory for deployment.

In the final stage, we implemented OpenAI's GPT-4o on the technical manual, further refining our approach based on the insights gained from previous experiments. The introduction of enhanced prompts and more sophisticated retrieval techniques marked a significant leap in performance. Experiment 3, which applied a new prompt on the latest two models(chunking by sections, and chunking by chapter), showed the most promise. The feedback confirmed that GPT-4o, combined with these refined prompts, was highly effective in generating accurate, detailed, and

contextually appropriate responses. The model's ability to parse complex documents and retrieve precise answers was significantly improved, particularly when focusing on specific chapters of the manual.

Both the lastest models, with different chunkings were performative, but The most noteworthy outcome was from the experiment where we split the entire document by sections, and with the use of gpt-4o. This approach proved to be the most performative, as it allowed the model to focus on relevant sections, greatly enhancing the accuracy and relevance of its responses. This method provided clear benefits in terms of precision and reliability. Given its superior performance, this model with enhanced prompts will be the one we deploy on the company website. It offers the best balance of accuracy, contextual understanding, and user satisfaction, making it the optimal choice for real-world applications in technical question-answering tasks related to reverse osmosis systems.

4.5 GraphRag

4.5.1 Introduction

In this section, we explore the implementation of GraphRAG, an advanced technique that integrates graph-based data structures with Retrieval-Augmented Generation (RAG) models. GraphRAG represents a significant evolution in the way we retrieve and utilize information, particularly in complex domains like reverse osmosis (RO) systems. Traditional RAG models focus primarily on extracting information from linear or hierarchical datasets, but they often fall short when dealing with intricate relationships and dependencies between data points. GraphRAG, on the other hand, leverages the power of knowledge graphs to better understand and represent these relationships, allowing for more accurate and contextually relevant responses.

The decision to implement GraphRAG was driven by the need to overcome the limitations of previous models, such as Llama3 and GPT-4o, which, despite their advanced capabilities, struggled with certain aspects of information retrieval in highly specialized technical domains. By integrating graph-based data, we aim to enhance the model's ability to navigate complex documents and provide more precise answers to technical questions. The primary objective of this implementation is to improve the model's performance in terms of accuracy, relevance, and contextual understanding, making it better suited for deployment in real-world applications within the RO industry.

4.5.2 Data Preparation for GraphRAG

The data preprocessing for our GraphRAG implementation was centered around a single, highly detailed document: the markdown file of the DuPont Technical Manual. To efficiently process this document and create a knowledge graph, we utilized code provided by Microsoft specifically designed for GraphRAG. Microsoft uses LLMs to generate and extract entities, communities, and relationships from the markdown file, constructing a knowledge graph that reflects the complex connections within the technical content.

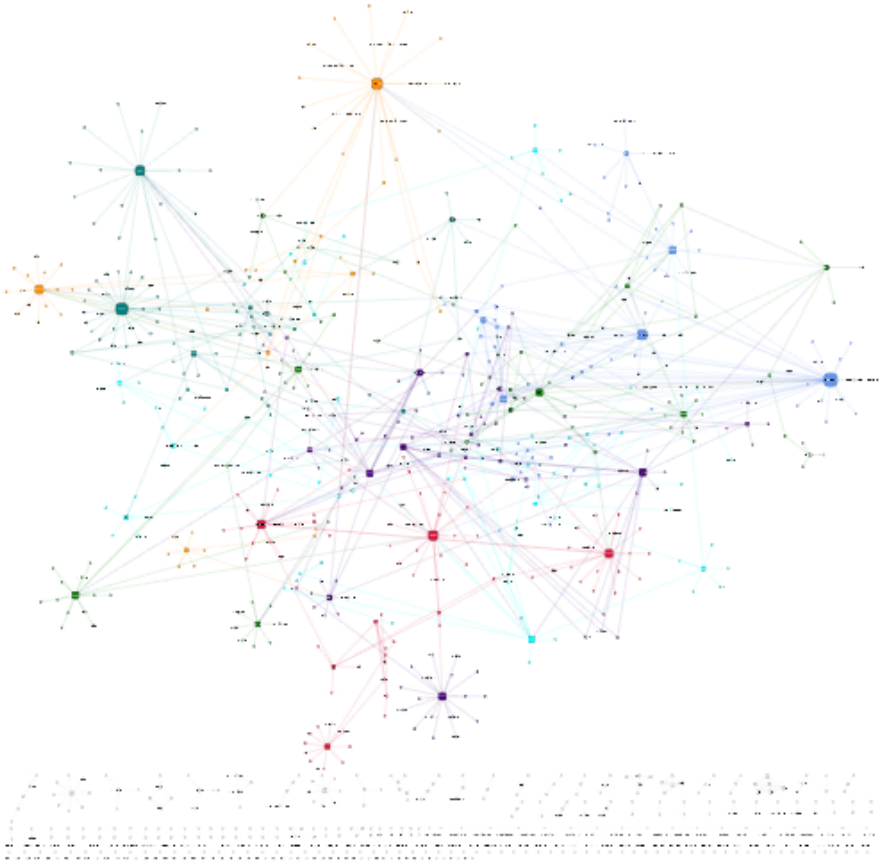

Figure 4.4: An overview of the generated graph using yfiles_jupyter_graphs library from python.

Entities, communities, and relationships are then stored in Parquet files format, which is optimized for large-scale data storage and retrieval. The use of Parquet files ensures that the data remains easily accessible and efficiently manageable, allowing for quick retrieval during the question-answering process.

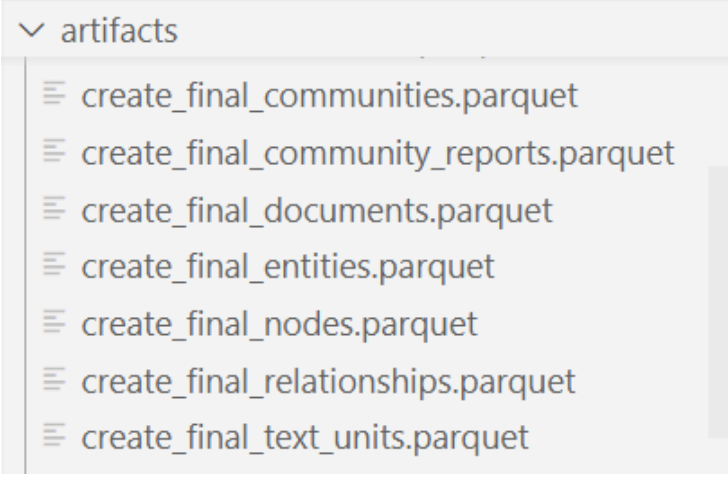

Figure 4.5: Data files generated by an LLM with predefined prompts by Microsoft.

The figure above shows a list of Parquet files generated during the data preprocessing stage for the GraphRAG implementation. Here's a brief overview of each file:

- **create_final_communities.parquet:** Contains information about various communities identified within the data. Communities represent clusters of related entities or concepts within the knowledge graph.
- **create_final_community_reports.parquet:** Includes reports or summaries related to the identified communities. These reports detail the characteristics of each community within the knowledge graph.
- **create_final_documents.parquet:** Stores information about the documents that were processed, in our case, it's only one document
- **create_final_entities.parquet:** Contains the entities extracted from the document, such as technical terms which are crucial for constructing the knowledge graph.
- **create_final_nodes.parquet:** Represents the nodes within the knowledge graph, each corresponding to an entity or concept extracted from the document.
- **create_fina_relationships.parquet:** Describes the relationships between the nodes (entities) in the graph, outlining how different concepts are interconnected.
- **create_final_text_units.parquet:** Holds the text units that were extracted and processed from the document, which are the basis for generating the entities and relationships.

4.5.3 Experiments

Experiment 1

Model name	gpt-4o
Open source	No
Embedding	text-embedding-3-small
Data	2021 - DuPont FilmTec™ RO Technical Manual
Prompt_template	—Role— As an assistant for question-answering tasks specializing in Reverse Osmosis (RO) Water Treatment Systems, responding to questions about data in the tables provided. —Goal— Your goal is to provide accurate, concise. If you don't know the answer, just say so. Do not make anything up, or add information that is not answering directly the user's question Ensure your answers are direct, clear, logical, and tailored to the user's question, aiming to solve their problem effectively. Do not include information where the supporting evidence for it is not provided. —Target response length and format— response_type= 'multiple paragraphs' —Data tables— context_data
LLM Temperature	0

Table 4.4: Configuration details for the gpt-4o model and GraphRag used on the 2021 DuPont FilmTec™ RO Technical Manual.

The table above describes the configuration for the first experiment using GPT-4o. We got inspired by the original Microsoft system prompt, and with a few modifications, we adjusted it for our purpose for more accurate responses based strictly on the provided context. The use of the "text-embedding-3-small" embedding ensures efficient text representation, while the temperature setting of 0 guarantees deterministic outputs, minimizing the risk of fabrication or irrelevant information.

Feedback and discussion

> Aquadviser's feedback: in general the majority of the answers are good and have good details but too much details, and the conclusion in every answer, so it may not be practical for an operator to confuse them with all the non-necessary information and conclusion at the end of every answer.

According to the evaluation, while the model provided good and useful details, the answers were often broader than necessary. This could lead to confusion for operators who need straightforward and precise information. Some answers also failed to pick up the correct information from the data, highlighting the need for further refinement in focusing the model on the most relevant details. Despite these issues, the overall feedback was that the model generally performed well, but there is room for improvement in tailoring the responses more closely to the specific needs of the end-users.

Experiment 2

In this experiment, we observed that the variable response_type within the system prompt had a significant impact on the model's outputs. Initially, this parameter was set to generate responses in 'multiple paragraphs' by default, which sometimes resulted in answers that were more detailed than necessary. To address this, we decided to modify the respons_type to "single straight to the point paragraph." This change aimed to produce more concise, focused, and direct answers, aligning better with the need for straightforward information, especially in the context of technical queries related to Reverse Osmosis (RO) systems. There was no change in the parameters, except in the response_type variable.

Feedback and discussion

> Aquadviser's feedback: The answers are accurate and direct. If the answers to the questions are available in the DuPont technical manual, the model provides them precisely. If the information is not in the manual, the model correctly states that it is not available in the provided document. However, for answer number 6, the model did not provide an answer and incorrectly stated that the document does not specify the requested information, even though this information actually exists in the DuPont technical manual. This might be because the answer is in a table that the model could not extract. Additionally, answer 12 seems to be incorrect (hallucination), as it mentions "membrane... for low-salinity or cold seawater," which is not accurate. Answer 13 is also lacking in detailed insights.

The feedback indicates that the model successfully identified when information was missing from the document and refrained from making up answers, which is commendable. However,

there were instances of hallucinations, where the model provided inaccurate information, and in another question, where the response lacked depth. These observations suggest that while the model is largely reliable, improvements are needed in handling tabular data and ensuring accuracy to avoid hallucinations.

Experiment 3

In the final experiment, we aimed to address the feedback from Aquadviser by enhancing the depth and detail of the model's responses. To achieve this, we modified the response_type in the system prompt to instruct the model to generate a single, straight-to-the-point paragraph for the initial answer. Additionally, we added a new section titled "Additional information," where the model was prompted to provide more comprehensive details that the user might find useful. This adjustment was intended to not only maintain the accuracy and clarity of the primary response but also to offer richer, more informative content that aligns with the user's needs, thus improving the overall quality of the answers provided.

response_type="single straight to the point paragraph, add a section called: Additional information, that should be longer, in which you give more details that the user might need"

Feedback and discussion

> Aquadviser's feedback: 'The model in general is very concise and provides the clear cut info in the response and provides valuable additional info at the end. However, sometimes (A10) it repeated the response info exactly the same in the additional info. (Maybe this can be fine-tuned by prompting). Maybe the concept of saying The data provided does not specify the exact percentage of urea rejection by FilmTec BW30 RO membranes. is very true for suppliers like DuPont or Toray or any membrane suppliers because it means according to their published data there is no valid info about and hence, it can contact you directly with the suppliers such as DuPont by providing the email and the suggested tailored question (good thinking to discuss with DuPont and Mann Hummel).'

Overall, this model performed better than the previous ones. It consistently provided accurate, concise answers and effectively added valuable additional information when prompted. Despite some minor redundancies, the enhanced depth of the responses made this model the most effective and reliable of all the versions tested, demonstrating that the adjustments made significantly improved its performance.

Cost

Graphrag is known for being very expensive in terms of usage, especially in the indexing phase, and that is because to generate the knowledge graph, we use a LLM. For this reason, we wanted

to see how expensive was it to index the entire document of DuPont's technical manual

- **input tokens** : 913 790
- **input tokens price in US Dollars** : 3.74 $
- **output tokens** : 249 728
- **input tokens price in US Dollars** : 4.5 $
- **Total tokens** : 1 163 518
- **Total tokens price in US Dollars** : 8.31 $

4.5.4 Conclusion

The implementation of GraphRAG has shown significant potential in enhancing information retrieval for complex domains like reverse osmosis (RO) systems. By leveraging knowledge graphs, GraphRAG has improved the accuracy and relevance of responses, especially when dealing with technical documents such as the DuPont manual. The system's ability to understand and structure relationships between entities has proven useful in providing more precise and contextually aware answers. As we continue refining the model and integrating more data sources, GraphRAG will play a pivotal role in improving how AI assists in troubleshooting and decision-making for RO systems.

4.6 WebApp

The Aquadviser web application is designed to provide users with an easy-to-use interface for interacting with a conversational AI agent. The system uses several technologies and services to manage the website and the AI functions. Here's how it works:

- **Front-End Interface:** Users access the web application via the Aquadviser website, found at `https://www.aquadviser.com`. The application itself is hosted at `https://www.aquadviser.com/app`.
- **Backend Infrastructure:** The backend is hosted on Azure App Service, ensuring scalability and strong performance for the entire website. The platform is developed using FastAPI, which manages various sections of the website. For the conversational AI applications, we use Chainlit, which enables smooth communication between users and the intelligent agent.

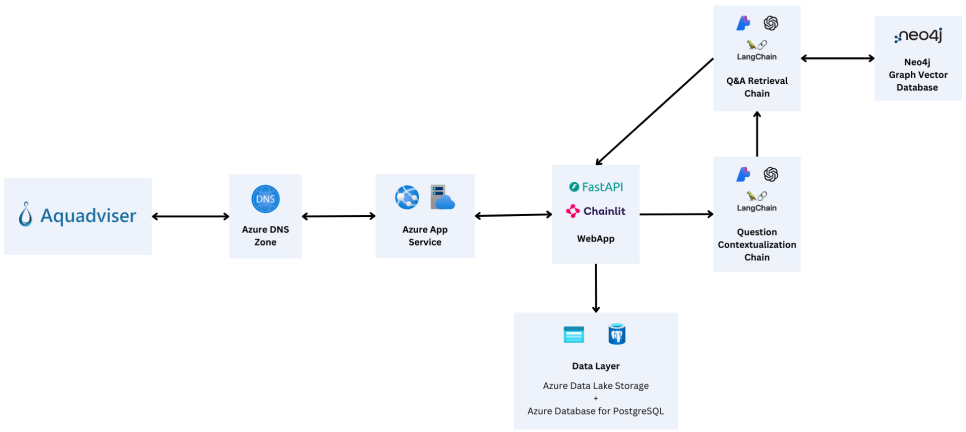

Figure 4.6: Web Application Architecture.

- **AI Agent (LangGraph):** At the core of the system an AI agent, is implemented using a LangGraph framework. This LangGraph Agent consists of the following nodes:
  - **Contextualization Chain:** This node improves the user's question by considering the context from their chat history, and reformulates the question when necessary.
  - **Q&A Retrieval Chain:** Once the query is contextualized, this node retrieves relevant information from the Neo4j Database to provide accurate answers to the questions.
- **Data Layer:** The application's data is stored in two primary sources:
  - **Azure Data Lake Storage** for managing large files.
  - **Azure Database for PostgreSQL** to handle structured data, particularly user related information, such as chat history and application-specific data.

This architecture ensures that the conversational AI agent can effectively respond to user queries by contextualizing them and retrieving the most relevant information. The modularity and scalability of the system also allow for future enhancements and modifications to the AI workflows.

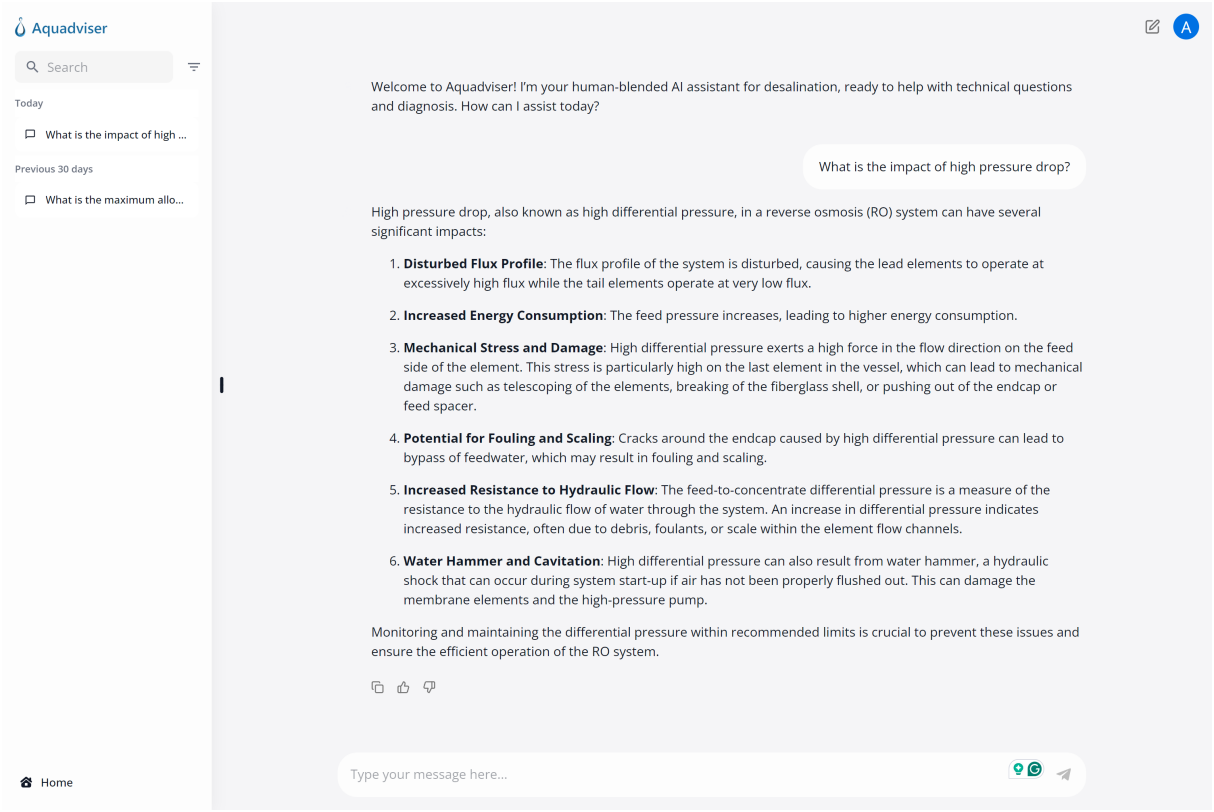

Figure 4.7: Web Application User Interface.

General Conclusion

This report presents an extensive study on the integration of Artificial Intelligence (AI) techniques, specifically focusing on the application of machine learning models and Large Language Models (LLMs), to improve predictive maintenance and troubleshooting in reverse osmosis (RO) systems. Starting from understanding the principles of reverse osmosis and identifying common challenges such as biofouling, scaling, and energy inefficiencies, the report delves into the use of advanced technologies like Retrieval-Augmented Generation (RAG) and GraphRAG to enhance the accuracy and relevance of technical support in RO systems.

A series of experiments were conducted using various LLMs—beginning with Llama3 and advancing to more sophisticated models like GPT3 and GPT-4o. Each experiment aimed to refine the model's ability to provide precise, contextually accurate answers by leveraging data from technical manuals, Q&A files, and other resources. We have discovered through These experiments, that continuous model fine-tuning and prompt engineering are a crucial part of achieving better results in LLMs. We also explored state-of-the-art methods such as GraphRAG, which was difficult due to the limited resources of the approach, however, we eventually managed to showcase how implementing graphs in RAG systems can have a significant potential in organizing and retrieving information more effectively.

Several limitations emerged throughout the study. The availability and quality of data were key challenges, with some models performing suboptimally due to the lack of comprehensive datasets. Moreover, while sequential models like LSTM were considered for predictive maintenance, the current dataset was insufficient to fully explore their potential, which hindered the ability to anticipate problems with high temporal accuracy. The integration of additional documents and more diverse data sources would further strengthen the model's ability to provide reliable, detailed responses across a broader range of technical scenarios.

Looking ahead, future perspectives include expanding the dataset, particularly with the addition of technical manuals, case studies, and operational data from real-world RO systems. We also plan to further optimize graph-based models like GraphRAG to enhance their ability to navigate complex information and improve decision-making capabilities, through enhancing the prompt design to ensure even more accurate responses. The development of more sophisticated prompts and algorithms tailored to specific RO-related tasks is crucial for achieving even greater precision

and operational efficiency. Overall, this work showcases the immense potential of AI-driven solutions in addressing the pressing challenges in water treatment systems, contributing to more intelligent, efficient, and sustainable operations within the industry.

Bibliography

[1] Shashikant Kumar, A Choudhury, Shilpa Saini, and Sanjeev S Katoch. A short review on process and applications of reverse osmosis. *International Journal of Scientific Engineering and Technology*, 2:1150–1156, 2013.

[2] Jozef M. Pacyna, Elisabeth G. Pacyna, Fred Steenhuisen, and Simon Wilson. Global atmospheric mercury assessment: Sources, emissions, and transport. *Environmental Science Technology*, 40(22):8054–8061, 2006.

[3] M. Jafari et al. Cost of fouling in full-scale reverse osmosis and nanofiltration installations in the netherlands. *Desalination*, 500:114865, 2021.

[4] M. B. Abid, R. A. Wahab, M. A. Salam, I. A. Moujdin, and L. Gzara. Desalination technologies, membrane distillation, and electrospinning, an overview. *Heliyon*, 9:e12810, 2023.

[5] https://www.gantt.com/.

[6] Complete Water. The history of reverse osmosis.

[7] Phillip. Elimelech, Menachem and William A. The future of seawater desalination: energy, technology, and the environment. *Science*, 333(6043):712–717, 2011.

[8] E.M.V. Hoek, T.M. Weigand, and A. Edalat. Reverse osmosis membrane biofouling: causes, consequences and countermeasures. *npj Clean Water*, 5:45, 2022.

[9] HB. Wang, YH. Wu, WL. Wang, et al. Biofouling characteristics of reverse osmosis membranes by disinfection-residual-bacteria post seven water disinfection techniques. *npj Clean Water*, 6:24, 2023.

[10] I. Shahonya, F. Nangolo, M. Erinosho, and E. Angula. Scaling and fouling of reverse osmosis (ro) membrane: Technical review. In M. Awang and S.S. Emamian, editors, *Advances in Material Science and Engineering*, Lecture Notes in Mechanical Engineering. Springer, Singapore, 2021.

[11] Çağla Odabaşı, Pelin Dologlu, Fatih Gülmez, Gizem Kuşoğlu, and Ömer Çağlar. Investigation of the factors affecting reverse osmosis membrane performance using machine-learning techniques. *Computers Chemical Engineering*, 159:107669, 2022.

[12] Yun Teng and How Yong Ng. Prediction of reverse osmosis membrane fouling in water reuse by integrated adsorption and data-driven models. *Desalination*, 576:117353, 2024.

[13] W. John Hutchins. The georgetown-ibm experiment demonstrated in january 1954. In *Conference of the Association for Machine Translation in the Americas*, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[14] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[15] Sepp Hochreiter. Long short-term memory. *Neural Computation*, 1997.

[16] Tomas Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[18] Meta AI. Meta ai blog: Meta llama 3. `https://ai.meta.com/blog/meta-llama-3/`, 2024. Accessed: 2024-09-09.

[19] Abhimanyu Dubey et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[20] OpenAI. Gpt-4 openai system card, 2023. Accessed: 2024-09-09.

[21] Wielded Blog. Gpt-4o benchmark: Detailed comparison with claude and gemini, 2024. Accessed: 2024-09-09.

[22] Erwin Rimban. Challenges and limitations of chatgpt and other large language models. *SSRN Electronic Journal*, page n. pag., 2023.

[23] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Sebastian Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

[24] Prompt Engineering Guide. Retrieval-augmented generation (rag). Accessed: 2024-09-09.

[25] Google Cloud. Retrieval-augmented generation use case. Accessed: 2024-09-09.

[26] Inc. Amazon Web Services. What is retrieval-augmented generation (rag)?, 2024. Accessed: 2024-09-09.

[27] NVIDIA. What is retrieval-augmented generation (rag)?, 2023. Accessed: 2024-09-09.

[28] Databricks. Retrieval-augmented generation (rag) - what is rag?, 2023. Accessed: 2024-09-09.

[29] Darren Edge et al. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

[30] Tianqi Chen and Carlos Guestrin. Xgboost documentation, 2016. Accessed: 2024-09-09.

[31] IBM. Random forest, 2023. Accessed: 2024-09-09.

[32] IBM. What is linear regression?, 2024. Accessed: 2024-09-09.

[33] GeeksforGeeks. Selenium python tutorial, 2024. Accessed: 2024-09-09.

Summary

This project explores the integration of Artificial Intelligence (AI) techniques to enhance the performance of reverse osmosis (RO) systems used in water treatment, particularly in desalination. By applying machine learning models for predictive maintenance and utilizing Large Language Models (LLMs) for troubleshooting, we aim to optimize system performance and reduce operational costs. The study involved multiple experiments using advanced models such as Llama3, GPT-3, and GPT-4o, combined with Retrieval-Augmented Generation (RAG) and GraphRAG for more effective information retrieval. Despite challenges like data availability and the need for further optimization of models, significant progress was made in improving the accuracy of predictive maintenance and troubleshooting. The project demonstrates the potential of AI to address critical challenges in water treatment systems, offering a more intelligent, efficient, and sustainable approach to managing RO operations.

Résumé

Ce projet explore l'intégration de techniques d'intelligence artificielle (IA) pour améliorer les performances des systèmes d'osmose inverse (RO) utilisés dans le traitement de l'eau, en particulier dans la désalinisation. En appliquant des modèles d'apprentissage automatique pour la maintenance prédictive et en utilisant des modèles de langage (LLM) pour le dépannage, nous visons à optimiser les performances des systèmes et à réduire les coûts d'exploitation. L'étude a impliqué plusieurs expériences avec des modèles avancés tels que Llama3, GPT-3 et GPT-4o, combinés avec la génération augmentée par la récupération (RAG) et GraphRAG pour un meilleur accès à l'information. Malgré des défis comme la disponibilité des données et la nécessité d'optimiser davantage les modèles, des progrès significatifs ont été réalisés dans l'amélioration de la maintenance prédictive et du dépannage. Ce projet démontre le potentiel de l'IA pour répondre aux défis critiques des systèmes de traitement de l'eau, offrant une approche plus intelligente, efficace et durable pour la gestion des opérations de RO.