Implementation of a Chatbot for Downtime Data Analysis Using Retrieval-Augmented Generation

Degree programme: Master of Science in Engineering
Specialisation: Data Science
Thesis advisor: Prof. Dr. Jürgen Vogel
Expert: Prof. Dr. Daniele Puccinelli (SUPSI)
Industrial partner: Hoffmann Neopac AG, Oberdiessbach

Manufacturing companies collect large amounts of production data, yet valuable insights often remain hidden in complex systems and free-text records. This project demonstrates how a chatbot leveraging Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) can facilitate natural language access to downtime data, reduce analytical effort, and support data-driven decision-making in an industrial environment.

Introduction

Manufacturing Execution Systems (MES) generate large volumes of structured and unstructured data that are essential for analyzing production performance and machine downtimes. At Hoffmann Neopac AG, a leading producer of plastic and laminated tubes, analyzing this data remains time-consuming and requires specialized knowledge. To address this challenge, this project develops a chatbot prototype that applies LLMs in combination with RAG to provide efficient and user-friendly access to MES downtime insights.

Objectives

This master's thesis builds on previous work and advances an existing chatbot prototype. The project pursued three objectives: evaluating alternative LLMs, establishing a test environment for user feedback, and improving the user experience through targeted interface enhancements.

Methods

A structured methodology combining data preparation, system extension, and evaluation was applied. The chatbot was implemented using a modular, containerized architecture with a Neo4j graph database, a FastAPI-based backend, a feedback API, and a Streamlit frontend.
The backend supported both serverless and locally hosted LLMs to allow consistent comparison under identical conditions. Model performance was evaluated using accuracy and RAG-specific metrics, as well as system indicators such as inference time and cost. The system was deployed in a containerized Azure environment for user testing, followed by a System Usability Scale (SUS) survey to assess usability and user acceptance after a four-week test phase.

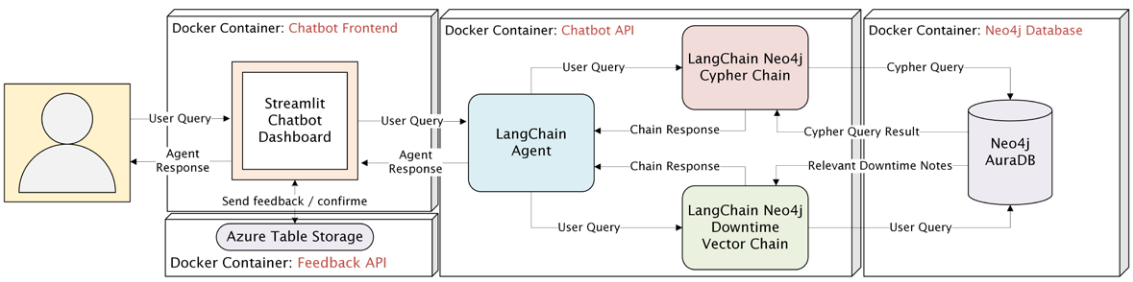
Results

The results show that serverless models achieved higher accuracy and more stable RAG-specific scores. Lightweight models such as GPT-4o-mini provided a strong balance between answer quality, inference time, and cost and were therefore selected for user testing.
The chatbot achieved an average SUS score of 70.83, slightly above the common benchmark of around 68, indicating good usability without the need for extensive technical training. User testing further revealed gaps between technically correct answers and perceived usefulness, especially for unclear or domain-specific queries, highlighting the importance of transparency and usability alongside model performance.

Conclusion

This work shows that combining LLMs with RAG enables efficient and user-friendly access to MES downtime data in an industrial context. The findings also emphasize that system design, usability, and continuous user feedback are key factors for successful productization.

Kilian Schürch
kilian.schuerch@gmail.com



Flowchart of LLM RAG Chatbot for MES Downtime Data