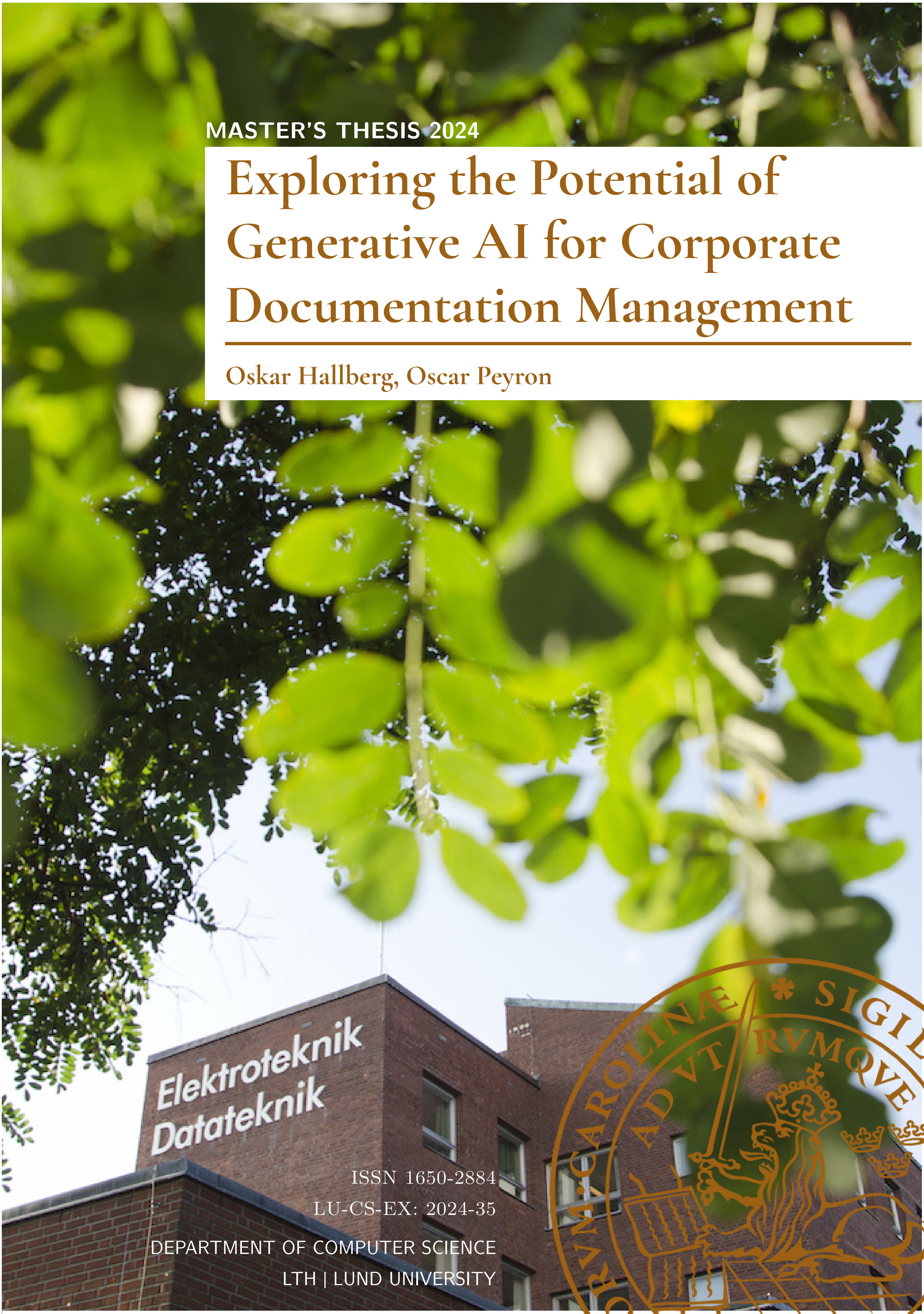
MASTER'S THESIS 2024

Exploring the Potential of Generative AI for Corporate Documentation Management

Oskar Hallberg, Oscar Peyron



ISSN 1650-2884
LU-CS-EX: 2024-35
DEPARTMENT OF COMPUTER SCIENCE
LTH | LUND UNIVERSITY

EXAMENSARBETE
Datavetenskap

LU-CS-EX: 2024-35

Exploring the Potential of Generative AI for Corporate Documentation Management

Utforskande av potentialen för generativ AI vid hantering av företagsdokumentation

Oskar Hallberg, Oscar Peyron

Exploring the Potential of Generative AI for Corporate Documentation Management

Oskar Hallberg
os1364ha-s@student.lu.se

Oscar Peyron
os4776pe-s@student.lu.se

June 20, 2024

Master's thesis work carried out at CS.

Supervisor: Elizabeth Bjarnason, elizabeth.bjarnason@cs.lth.se

Examiner: Alma Orucevic-Alagic, alma.orucevic-alagic@cs.lth.se

Abstract

Efficient and accurate documentation is an important element for maintaining effective workflows in corporations. This thesis investigates the integration of generative AI into documentation processes of a multinational corporation to address inefficiencies in manual documentation management. Using Design Science research complemented by Cross-Industry Standard Process for Data Mining (CRISP-DM), we developed and evaluated an AI-driven solution tailored to automate and enhance the creation of documentation. Through interviews and thematic analysis, we identified key challenges such as scalability, time consumption, inaccuracies, and resistance to technology adoption. Our solution employs fine-tuning, Retrieval-Augmented Generation, and prompt engineering to generate accurate and contextually relevant documents. The solution demonstrated improvements in documentation efficiency and quality while reducing manual errors. However, integration challenges and the need for continuous model training were noted. The findings suggest that while AI can improve documentation processes, ongoing adjustments and adaptations are essential for maintaining alignment with corporate standards and practices.

Keywords: Generative AI, Documentation Management, CRISP-DM, Design Science Research

Acknowledgements

We would like to thank the case company for their collaboration and support throughout this Master's thesis. Their willingness to provide access to their documentation processes and challenges has been a key component in the success of this project.

We are profoundly thankful to our supervisor, Elizabeth Bjarnason, for her invaluable guidance and insight throughout this research journey. Her constructive feedback, persistent support, and attention to the structure of our report have enhanced the quality of our work.

Special thanks are due to Martin and the development team, as well as Isabell and Noah. As our primary contacts, they provided essential support and insights that were crucial to the practical outcomes of our study.

Contents

Chapter 1

Introduction

Corporations are continuously faced with the imperative to enhance their products and operational processes to remain competitive in the business landscape. Communication and information management stand out as hurdles, requiring innovative solutions to ensure that relevant information is widely and accurately delivered across the organisation. This challenge highlights the necessity for large enterprises to adopt advanced digital technologies to improve their internal processes and maintain a competitive edge in the rapidly evolving business landscape.

The case company investigated in this thesis is a global industry leader that exemplifies the challenges faced by large corporations in managing complex and dynamic information flows. The company's vast and diverse array of documentation, including policies, guidelines, and product information, underscores the need for efficient tools to support its international workforce and operations.

The advent of artificial intelligence (AI) technology, particularly in its generative capacity, presents a promising solution to these challenges. Generative AI [17] has the potential to revolutionise the way corporations manage information by automating the creation and revision of documents. This technology not only has the potential of reducing manual labour but also to enhance the accuracy and accessibility of information, facilitating better decision-making and productivity across the organisation.

This thesis focuses on exploring the application of generative AI in automating documentation processes within the case company. By investigating the potential of AI to address the inefficiencies of manual documentation practices, this study aim to demonstrate how leveraging AI technologies can lead to more efficient and accurate corporate documentation.

1.1 Aims and Research Questions

Through our work, we seek to examining how generative AI can be applied to automate the generation of documentation at the case company. We aim to explore the potential of

AI technologies to create documentation automatically, thereby addressing the challenges posed by manual documentation processes. Although the hope is that such automation will enhance the overall process, our focus is not on evaluating the process improvement per se but rather on the capability of generative AI to generate necessary documentation efficiently and accurately. This exploration is rooted in understanding the current manual handling of documentation and developing as well as assessing an AI-based solution for the automated generation of documentation. The following research questions guide our investigation:

RQ1: What are the key challenges associated with manual documentation handling processes within a corporate setting?

RQ2: How can documentation be created automatically using generative AI?

RQ3: What is the quality and accuracy of documentation produced by a generative AI application?

These research questions aim to uncover the potential of generative AI in transforming documentation practices, highlighting both the opportunities and challenges of its implementation in a corporate environment similar to the case company. Through this investigation, the thesis aspires to contribute insights into leveraging AI for enhancing operational workflows and documentation management in a corporate setting.

The outcomes of this research are anticipated to be of value not only to the case company but also to other large enterprises facing similar documentation challenges. By demonstrating the practical use of generative AI in automating documentation processes, this study could serve as a benchmark for other corporations striving to enhance their documentation efficiency and accuracy.

1.2 General Approach

This thesis adopts the Design Science research methodology [23][40], complemented by the Cross-Industry Standard Process for Data Mining (CRISP-DM) [34], to systematically explore the use of generative AI for documentation generation at the case company. This approach is grounded in a combination of established knowledge from literature and an in-depth understanding of the specific context and challenges faced by the case company.

To address RQ1, we embark on an investigation of the case company's documentation challenges, utilising data collection methods such as document studies and interviews. This phase aims to gain an understanding of the current manual documentation processes and their associated inefficiencies, serving as the foundation for our subsequent design efforts.

For RQ2, the focus shifts to the design and implementation of a generative AI solution tailored to automate the generation of documentation. This involves the iterative development of an artefact – our proposed solution – based on insights derived from the interviews and the knowledge gained through a literature review.

RQ3 is addressed through the evaluation of the implemented solution within the case company's operational environment. This evaluation considers key factors such as the solution's effectiveness in generating accurate and compliant documentation.

Our adoption of the Design Science Research framework, enriched with the procedural approach of CRISP-DM, establishes a structured and iterative approach to designing, implementing, and evaluating the AI-based solution. This method not only facilitates the development of a practical solution to the identified documentation challenges but also enables us to assess its applicability and effectiveness in a real-world corporate setting.

1.3 Case Description

Our research was conducted at the case company which is a multinational corporation with a formidable global presence. The organisation operates in over 30 countries and contains a workforce of more than 170,000 individuals, specialising in a broad spectrum of retail services.

This large-scale operation requires a robust system for managing customer inquiries and providing accurate responses. The user flow shown in Figure 1.1 details how inquiries and content is currently managed. *Customers* come in contact with *customer support*, who may directly respond using their expertise or access the case company's knowledge platform. This platform is connected to a database that contains various documents which can be used by the customer support to provide a response to the customer. If the information is still lacking, customer support may raise the issue to *knowledge specialists* or another *co-worker* to seek further answers. For recurring similar requests, there is a convoluted process to create new documents which involves both knowledge specialists and *document writers*. When a document has been created and accepted, it is then added to the database to aid future support interactions.

The tasks of generating, managing, and processing documentation present several challenges. The volume of documents, ranging from topics such as frequently asked customer inquiries, internal process descriptions, policy documents, guidelines and detailed product information, requires resources for their creation, review, and maintenance. Moreover, ensuring that documents are accessible, follow company standards, and enables collaboration across departments and geographic locations adds another layer of complexity. As documentation cannot stay static, it needs to evolve and be updated as products, policies, and standards change. Manual documentation processes can be time-consuming and prone to errors. Thus, manual documentation management can lead to high costs and operational inefficiencies. For these reasons, automating documentation processes is appealing. However, shifting from manual to automated document management requires investment in technology and training.

Our case study examines the documentation processes at the case company, exploring how these can be refined and automated through the application of generative AI technologies. The primary stakeholders in this endeavour include:

- *Customer support:* Dependent on up to date, accessible and correct documentation in order to serve customer inquiries.

- *The knowledge documentation teams:* Charged with generating and revising corporate materials, these teams are poised to benefit directly from the efficiencies brought forth by AI-driven automation.

- *The general case company's employees:* Dependent on precise, timely, and readily accessible internal documentation to effectively fulfil their roles.

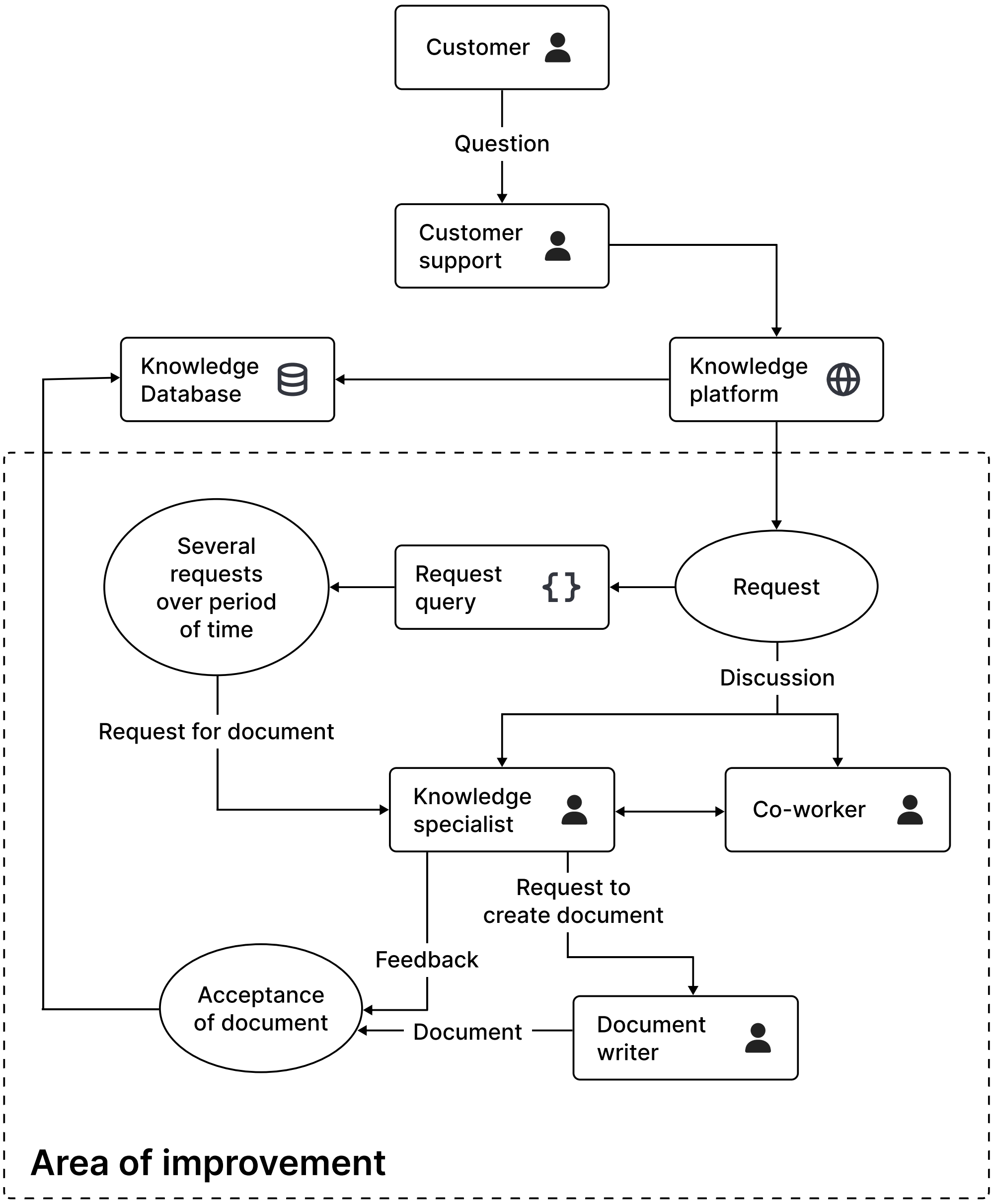


Figure 1.1: The current user flow for documentation processes within the case company.

1.4 Division of Work

Both authors engaged collaboratively in all aspects of this thesis, with the majority of the work conducted jointly. Nevertheless, specific tasks and sections were allocated based on individual strengths and focuses to optimise our workflow.

In terms of implementation, Oscar Peyron assumed the lead in designing and building the application. He focused on integrating LangChain [45] with the AI models and enhancing the system's capabilities through Retrieval Augmented Generation and search engine capabilities.

Oskar Hallberg took the lead in the technical enhancement of the AI models. He was primarily responsible for tuning the models to optimise performance, engaging in prompt engineering, data preparation, evaluation and testing.

Regarding user interaction, both individuals took charge of different aspects of the interview processes. Oscar Peyron led the focused interviews. Meanwhile, Oskar Hallberg took the lead in the conducted application evaluations.

The writing of the thesis was a parallel effort. Both authors actively contributed to drafting and revising the manuscript. Each chapter received equal attention from both authors, ensuring a cohesive and comprehensive presentation of our findings. However, certain sections saw a more concentrated effort from one author than the other: Oscar Peyron focused more on Background and Related Work (Section 2), Research Method (Section 3), as well as Result and Discussion for RQ1 (Section 4.1, 4.2, 5.1). Oskar Hallberg concentrated more on the Introduction (Section 1), Result and Discussion for RQ2 (Section 4.3, Section 5.2) as well as Result and Discussion for RQ3 (Section 3.5, 5.3).

Chapter 2

Background and Related Work

Our thesis is based on previous work in the areas of Artificial Intelligence (AI), Natural Language Processing (NLP) and Large-Language Models (LLMs) including frameworks for developing AI-based applications, relating to existing research on the implementation.

2.1 Artificial Intelligence (AI)

The concept of AI is described by John McCarthy, who coined the term in 1955, as "...that of making a machine behave in ways that would be called intelligent if a human were so behaving" [36].The goal of AI is to create systems that can perform tasks that would typically require human intelligence and actions. In an article discussing AI, Marr explained the goal of AI as "...to identify and solve tractable information processing problems" [33]. These tasks could include reasoning, learning, problem-solving, perception, and understanding natural language. Some argue that AI has become a collective term for several subfields which has led to AI being described more as an umbrella term [41]. One of these fields are Neural Networks which can be described as a specific type of AI model structured to mimic the way human brains operate to process information [1][19]. Examples of other subfields are Machine Learning (ML) and Natural Language Processing (NLP).

ML enables systems to learn from data and the surrounding environment in order to emulate human intelligence and improve performance such as effectiveness in processing data and making predictions [14]. NLP allows machines to understand and interpret human language, facilitating the human-computer interactions. Since this thesis relies on the quality of language and documentation, NLP will play an important role.

2.1.1 Natural Language Processing (NLP) and Large Language Models (LLM)

While NLP enables computers to understand and interact with human language, LLMs are the models that apply these principles. NLP combines computational linguistics-rule-based modelling of human language with statistical methods, machine learning, and more [9]. Chowdhary describes NLP as "...a collection of computational techniques for automatic analysis and representation of human languages, motivated by theory". These technologies enable computers to process and analyse large amounts of natural language data, from speech recognition, language translation, and sentiment analysis to more complex tasks like automatic summarising, relationship extraction, and topic segmentation.

A common type of model that utilises this is LLMs, which in itself is a subset of AI focusing on NLP and generation tasks [8]. The deployment of advanced LLMs is exemplified by the Generative Pre-trained Transformer (GPT) series, for instance the GPT-3 model [18]. These models are adept at mimicking human-like text generation, answering queries, and executing a broad spectrum of linguistic tasks with remarkable precision, thanks to their training on extensive on large text-based datasets.

LLMs have made progress in recent years but there are still many limitations. For example, contextual misinterpretation may occur when a model does not grasp or receive the correct context needed to be able to produce a correct result. Similarly, other limitations such as censorship, bias and misinformation are widely discussed topics in recent times [7]. Furthermore, the lack of interpretability makes it difficult to understand the reasoning behind a model's predictions [28].

LLMs have and will continue to push the boundaries of what is possible in the realms of NLP thanks in large to several techniques such as pre-training, Retrieval-Augmented Generation (RAG) and fine-tuning. These techniques have proven central when creating successful LLMs [15][43][5].

2.1.2 Conducted AI Approaches

Pre-training, RAG, fine-tuning and reinforcement learning are some of many important approaches in the development of AI models, especially in the context of NLP and LLMs. Together or individually, these techniques provide a powerful framework for developing effective NLP models by leveraging the benefits of large-scale data learning and retrieval with targeted specialisation.

In the phase of pre-training, models undergo initial training on large, diverse datasets to establish a broader understanding of language structures, nuances, and general knowledge [21]. This foundational training equips the models with a broad linguistic and contextual grasp. With the foundation laid it can be of interest to make the model more specialised on a specific domain or task which is where fine-tuning and RAG become relevant. RAG fetches information from a specified data source in real time and incorporates this into the response of the model while fine-tuning incorporates the additional knowledge into the model itself via training [4]. The process of fine-tuning is to adjust the pre-trained models to excel in specific tasks by training further on smaller, task-oriented datasets. The method of combining these activities enhances various aspects of model performance, including accuracy, efficiency in learning task-specific nuances, and adaptability to different domains or applications. This can be seen with models such as the previously mentioned GPT-series as well as Bidirectional Encoder Representations from Transformers (BERT), which after completion of pre-training, can be fine-tuned for tasks ranging from sentence classification to question answering [12].

While pre-training and fine-tuning is rather common, RAG is a rarely incorporated technique for public models since it is highly personalised and requires computational resources to implement effectively. The process of dynamically retrieving relevant information from data sources before generating a response means that RAG models necessitate access to vast, well-organised, and up-to-date knowledge bases [30]. This setup is both data and computationally intensive, involving indexing, searching algorithms, and storage solutions to facilitate rapid retrieval, which can be a barrier for widespread adoption, especially by smaller organisations or individual developers.

Reinforcement learning can enhance model performance by allowing it to iteratively improve its responses based on the outcomes of interactions [49][31]. This iterative process enables the model to adapt and optimise its behaviour over time, leading to more effective communication and problem-solving capabilities. While reinforcement learning offers potential for enhancing AI models, its implementation can be complex and resource-intensive. Training reinforcement learning agents often requires substantial computational resources and careful design of reward functions to ensure effective learning [16].

Nonetheless, these techniques have proven effective when implemented which has led to a growing incorporation of LLMs in different business aspects [26].

2.1.3 Adoption of AI and LLMs in Business Practices

The integration of AI and LLMs into business operations has marked a shift in how companies approach problem-solving, innovation, and customer engagement [46]. As these technologies continue to evolve, their adoption across various industries is not just a trend but a testament to their potential to transform traditional business models. In this section, we investigate the dynamics of AI and LLMs and their growing influence in the business world, where their usage provides a range of advantages, but also poses some challenges.

Usage and Advantages of AI and LLMs

There are a wide array of business functions that make use of AI and LLMs, thereby demonstrating the versatility and potential efficiency of these technologies [10]. AI and LLMs facilitate data-driven decision-making, providing businesses with insights derived from large datasets that human analysis could not feasibly process. One of the primary uses is in customer service, where chatbots and virtual assistants, powered by LLMs, provide 24/7 support, handling inquiries and resolving issues with strong accuracy and human-like understanding [24].

Patel and Trivedi came to the conclusion that the implementation of AI, Machine Learning and NLP in customer support will provide businesses with a greater opportunity to analyse data and therefore uncover useful and valuable insights such as customers loyalty [38].

Haleem et al. argues that in marketing, encompassing techniques like machine learning, empowers machines with human-like cognitive functions such as learning and reasoning [20].

This has enhanced personalised brand experiences, boosting user engagement and loyalty. Language-based AI tools have transformed marketing, serving as sales assistants, payment processors, and customer engagement managers. These tools simplify purchasing processes and learn from interactions to improve future experiences. AI also enables the personalising of content and optimises email marketing campaigns by analysing data to offer insights, making marketing strategies more effective and data-driven.

Challenges to Adopting AI

Despite the clear benefits, the adoption of AI and LLMs is not without its challenges [53][44]. One of the primary concerns is the ethical implications, including privacy issues and the potential for bias in decision-making processes. Ensuring that AI systems are fair, transparent, and respect user privacy requires ongoing effort and governance. Additionally, the integration of these technologies into existing business infrastructures can be complex and costly, necessitating investment in training and development to fully realise their potential.

Another hurdle is that LLMs can be prone to generate information or responses that are not grounded in the facts or data they were trained on. This phenomenon, often referred to as "hallucination," occurs when the model confidently produces outputs that are factually incorrect or entirely fabricated [27][48]. This issue can be particularly challenging in scenarios where accuracy and reliability of information are critical. Hallucinations in LLM outputs stem from several factors. One primary reason is in the nature of how these models are trained. LLMs learn to predict the next word in a sequence based on probabilities derived from the training data, without an inherent understanding of truth or factual accuracy. When faced with topics that are underrepresented in the training data or when generating content on complex subjects, the model might "fill in the gaps" with plausible but incorrect or nonsensical information.

The constant evolution of AI and LLM requires businesses to engage in continuous learning and adaptation to harness their full potential. This continuous need for updates and education can be daunting, making it challenging for companies to fully leverage the potential of these technologies. Recognising this, solutions like LangChain offer a way forward, providing a framework that simplifies the integration and utilisation of AI and LLMs.

2.1.4 LangChain

LangChain stands as a transformative framework designed to streamline the incorporation of AI and LLMs into the fabric of business operations. At its core, LangChain is built upon the philosophy of democratising the access and efficiency of language AI technologies for businesses.

LangChain provides an approach to address the challenges organisations face when integrating AI and large language models (LLMs) [37]. By streamlining the technical complexities of these technologies, LangChain offers a framework that helps businesses deploy AI solutions. This method allows companies to maximise the utility of LLMs in various applications [45]. Overall, LangChain's framework can help organisations harness the capabilities of advanced language models to drive innovation and improve their business outcomes.

Leveraging LangChain for Advanced Business Applications

LangChain provides a comprehensive toolkit designed to streamline the deployment and use of language models [25]. Its framework facilitates the development of applications that leverage the natural language processing capabilities of LLMs, from simple chatbots to complex analytical tools [45]. By abstracting the complexities involved in interfacing with LLMs, LangChain enables developers to focus on creating value-added features and services.

The toolkit is composed of libraries in both Python and JavaScript, making it accessible to a wide range of developers and application scenarios. Several modular components are included in the libraries. The following types of components are used in this thesis:

- *LLMs*: These are the components that handle the generation and understanding of natural language. LangChain supports integration with models such as the GPT-series.
- *Document Loaders*: These components are responsible for loading and preprocessing text data from various sources.
- *Document Transformers*: These transformers handle tasks such as extraction and splitting text into a format suitable for processing by LLMs.
- *Vector Stores*: These integrations facilitate communication with databases to store vector representations of documents.
- *Retrievers*: Retrievers are used to fetch relevant information from a dataset based on queries generated by the LLMs.
- *Agents*: These components enable the development and usage of specialised agents that can perform specific tasks.

Key Benefits of Implementing LangChain

One of the principal benefits of LangChain is its ability to mitigate the challenges associated with the "hallucination" phenomenon in LLMs. Through its handling and processing mechanisms, LangChain can help in reducing the occurrence of inaccurate or fabricated outputs by providing additional layers of validation and context. This ensures that the information generated by LLMs is more reliable and grounded in reality, which is crucial for businesses that depend on the accuracy of data-driven decisions.

Moreover, LangChain simplifies the continuous learning and adaptation process associated with the rapid evolution of AI technologies. Its framework is designed to be flexible and scalable, allowing businesses to update and expand their AI capabilities as new advancements emerge [37]. This reduces the technical barriers and resource investments required to stay at the forefront of AI technology, making it more feasible for companies of all sizes to leverage the benefits of LLMs.

2.1.5 AI-models Applied in this Thesis

AI models, particularly those designed for processing natural language, are typically built on neural networks composed of interconnected nodes or neurons [13]. These networks, often structured in layers, form the backbone of various machine learning applications. Each node in these layers connects through pathways that have associated weights, and each node itself possesses a bias. The training of these models involves adjusting these weights and biases according to the data they process, a method known as learning [6]. The learning objective is typically to minimise the discrepancies between the model's predictions and the actual outcomes, a process steered by what's called a loss function. This structure and method allow AI models to adapt and improve their accuracy and functionality over time.

Central to our thesis are two advanced models within the current landscape of NLP and LLMs, namely GPT-3.5 Turbo and GPT-4. These models are part of the GPT-series, building on the architecture established by GPT-3 [32]. The GPT-3 model is comprised of over 175 billion parameters where these parameters are essentially the weights and biases in the neural network that GPT-3 uses to generate text based on the input it receives [54][52]. This large amount of parameters enable GPT-3, and by extension GPT-3.5 Turbo and GPT-4, to generate complex and contextually relevant text. This capability makes them effective across a wide range of NLP tasks, often requiring minimal task-specific training.

Despite its advancements, GPT-3 is not devoid of challenges. The model is susceptible to biases inherent in the data on which it was trained [11]. This can lead to generation of content that provide misleading information, reflecting the biases present in the datasets used for training. Addressing this issue requires ongoing efforts to refine training methodologies and implement mechanisms to detect and correct biases in AI models. The capacity of such models to generate realistic and persuasive text raises questions about their potential misuse, including the creation of misleading information or impersonation. These challenges underscore the need for a balanced approach to harnessing the capabilities of AI, where innovation is matched with responsibility.

The GPT-3.5-Turbo Model

GPT-3.5 Turbo can be seen as an iteration built upon GPT-3.5 which in itself is built upon GPT-3, aiming to enhance performance, efficiency, and user experience. While it enhances the ability to generate human-like responses, it has also shown to comprise some task-solving ability [52]. This model is engineered to deliver responses with reduced latency, making it well-suited for applications that demand real-time interaction, such as conversational AI, where the immediacy of response is crucial. The improvements in speed are achieved without compromising the model's ability to understand nuances in language or generate high-quality and coherent text. The optimisations underpinning this model enable it to handle a larger volume of simultaneous requests.

The GPT-4 Model

GPT-4 is a continuation of the GPT-3.5 series with enhancements regarding complexity, contextual understanding, and application versatility [29]. This GPT-version is equipped with a vastly increased number of parameters, which enables a deeper comprehension of nuances and subtleties in language. GPT-4's architecture and training methodology have been refined to produce text that is not only coherent and contextually relevant but also capable of demonstrating an understanding of complex concepts and instructions.

The main differences between GPT-3.5 and GPT-4 are the models' amount of parameters

(170 Trillion for GPT-4 vs. 175 Billion for GPT-3.5) [29]. The size of the context length, GPT-4 have an ability to process longer strings of text which allows for a more nuanced understanding of the context. GPT-4 also introduces a multimodality feature that expands its application scope [2]. Unlike GPT-3.5, which primarily processes and generates text, GPT-4 can also interpret and generate responses based on both text and images. This multimodal capability opens new avenues for different type of AI applications.

2.2 Frameworks for developing Data-Driven Applications

There are various frameworks and methodologies for creating, deploying, and managing the development of AI-based or other types data-driven applications. This is done in order to aid in the task of processing large amount of data in corporate IT strategies [47]. Common frameworks, including SEMMA (Sample, Explore, Modify, Model, and Assess), KDD (Knowledge Discovery in Databases), and CRISP-DM (Cross Industry Standard Process for Data Mining), are tools for structured data analysis and mining [3][42]. These frameworks all share the goal of extracting insights and value from complex datasets but differ in methodologies, processes, and focus areas, catering to various project needs and organisational structures [3]. While SEMMA and KDD have a are more technically oriented, CRISP-DM has a strong focus on aligning data mining with business objectives. Because of this reason, as well as CRISP-DM's more adaptable and flexible nature, it was chosen as the framework for this thesis.

2.2.1 Cross Industry Standard Process for Data Mining (CRISP-DM)

The CRISP-DM framework is used for developing data-driven software, including those focused on the realm of Generative AI. CRISP-DM offers a structured approach that is both flexible and adaptable, making it suitable for the dynamic and exploratory nature of Generative AI endeavours. Wirth et al. describes the goal of the model as "...CRISP-DM process model aims to make large data mining projects, less costly, more reliable, more repeatable, more manageable, and faster" [50].

CRISP-DM facilitates the structured development of data-driven applications by guiding teams through six stages which includes *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation* and *Deployment* (see Figure 2.1). The process starts with the Business Understanding phase, where the focus is on comprehending the project's objectives and requirements from a business standpoint. This includes considering constraints, resources, risks, and assumptions, which informs the creation of a strategic plan to meet project goals. Following this, the Data Understanding phase involves initial data collection and various activities to familiarise with the data, such as analyses and quality checks. This leads into the Data Preparation phase, where the final dataset is constructed from the initial raw data. Key tasks here include removing missing values, treating outliers, and scaling the data. Next, in the Modeling phase, various techniques are selected and applied with parameters tuned to optimal values, aligning with the business requirements identified earlier. The subsequent Evaluation phase involves a thorough assessment of the model and a review of the steps taken during model construction to ensure alignment with business objectives. A crucial part of this phase is to ensure no significant business issue has been overlooked. Finally, the Deployment phase involves creating, testing, and deploying the model in real-time. This may include integrating the model into existing systems, launching new services, or making the outputs available to the intended audience, completing the cycle from understanding to action.

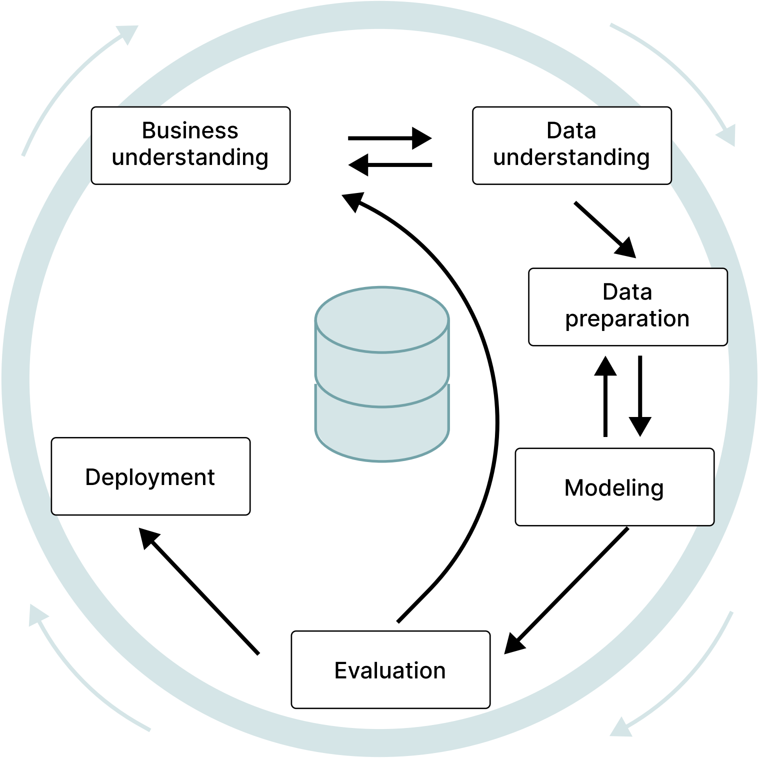


Figure 2.1: An overview of the CRISP DM Framework.

2.3 Documentation Evaluation

Evaluating generated documentation is a process aimed at ensuring the material meets the needs and expectations of its intended audience. This involves assessing several key aspects, such as clarity, understandability, accuracy and usability.

Evaluating documentation produced by AI presents unique challenges, particularly when relying on data-driven analysis methods. These methods, while effective in quantitative assessment, may struggle to fully capture the nuances of AI-generated documents. This is because AI documentation can vary widely in terms of style, context, and the intricacies of language, making standardised metrics less effective. Data-driven approaches primarily focus on technical aspects such as the density of keywords, readability scores, or error rates, which do not fully encompass the document's quality or its ability to meet user needs. Common evaluation techniques for text, such as BLEU (Bilingual Evaluation Understudy), often face challenges in producing definitive conclusions [22][39]. This limitation stems from their reliance on surface-level comparisons between the generated text and a set of reference texts. Consequently, these methods may not adequately capture the semantic accuracy or the contextual relevance of the generated content.

In contrast, human evaluation introduces a subjective dimension to the evaluation process and a more suitable method for assessing documentation. Human reviewers are better equipped to interpret the subtleties and context-specific elements that AI might introduce, which often elude algorithmic analysis. For example, they can determine whether the document is not only factually accurate but also contextually appropriate and engaging for the intended audience. Through user feedback, the voices of those who navigate the documentation daily are heard, revealing the strengths and areas for improvement from the user's perspective.

Previous work on this topic strengthens several claims that are useful for this thesis, namely how the quality of data plays a central point, the variance of data-driven evaluation methods and the comparison of human evaluation. Wiseman et al. performed an investigation by introducing a large-scale corpus of data records of sports game data in conjunction with descriptive documents and a series of extractive evaluation models [51]. This investigation shows the complexity of generating documentation that fits or outperforms the golden standard set by humans.

In this thesis, we analyse evaluation data as well as feedback from users to assess the effectiveness of AI-generated documentation. We employ both evaluation methods to offer perspectives on the strengths and weaknesses of the documentation. Integrating these insights allows us to develop an understanding of the documentation's performance in terms of various metrics and user experiences.

Chapter 3
Research Method

We applied a Design Science Research [23][40] approach combined with CRISP-DM [34] to explore the problems in manual documentation management at the case company and to design, implement, and evaluate an AI-based solution for that context. An overview of this process, which was inspired by an already published thesis written by Noah Mayerhofer and Sandra Nyström [35], is illustrated in Figure 3.1.

It was essential to acquire a thorough understanding of the organisational context and the specific documentation needs of the case company to articulate clear objectives for our AI-driven solution, which in turn guided the design and development of our application. Through this approach, we aimed to ensure that our solution was not only technically sound but also closely aligned with the strategic business requirements and documentation standards of the organisation.

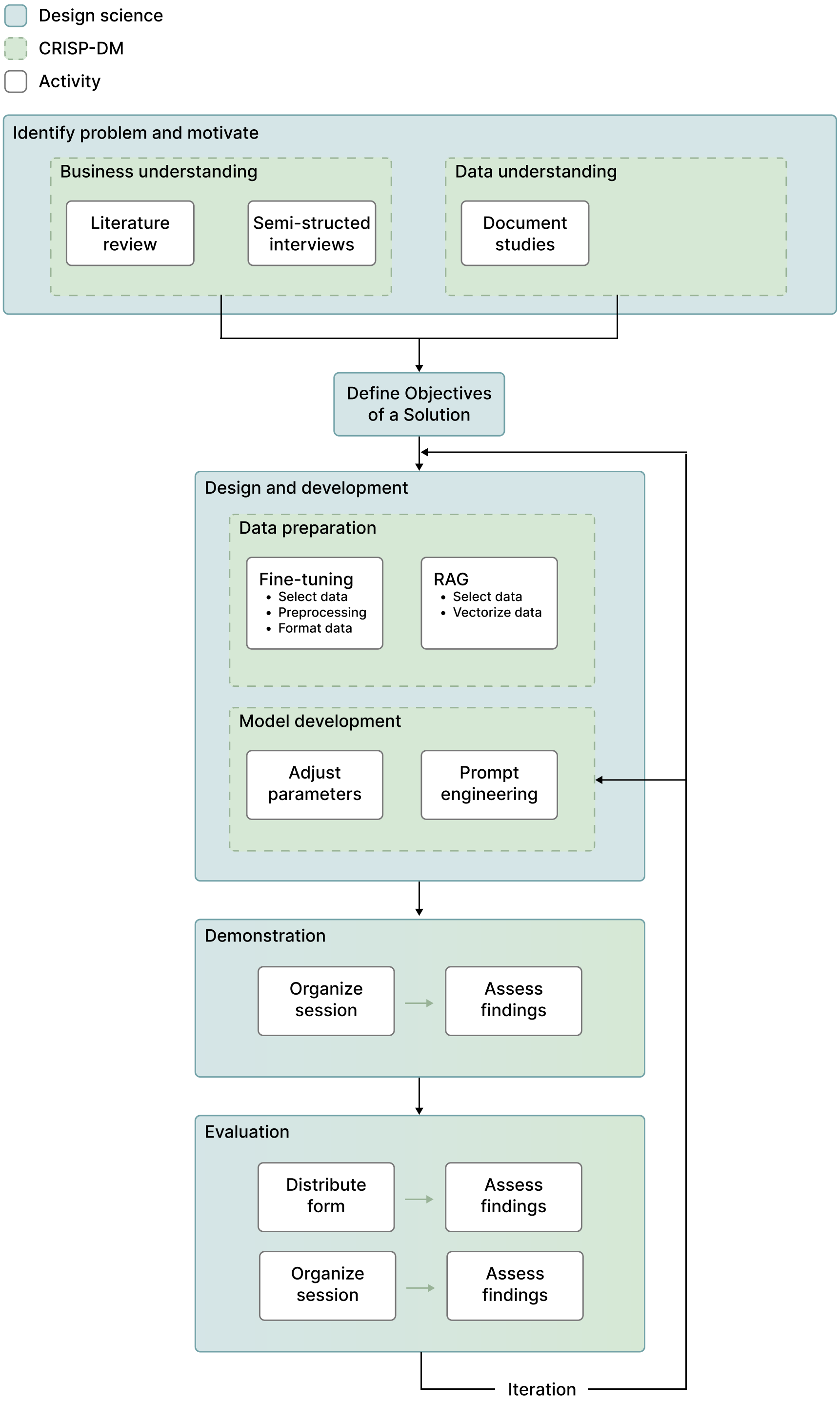


Figure 3.1: Overview of the thesis process inspired by Noah Mayerhofer and Sandra Nyström [35].

3.1 Identify Problem and Motivate

The initial step involved identifying and justifying the core problem with manual documentation processes. This step required a comprehensive understanding of the case company's business domain, the specific challenges associated with manual documentation processes, and the data used within the business domain.

3.1.1 Business Understanding

We conducted a literature review on generative AI and engaged key stakeholders through semi-structured interviews to understand the operational needs and expectations within the case company.

Literature Review

The literature review help us to grasp the broader business implications of AI technologies and the specific challenges and opportunities associated with their implementation in documentation workflows. Our objective was to uncover how generative AI technologies are being applied to enhance operational efficiencies, improve document accuracy, and streamline information dissemination within corporations.

For collecting relevant literature, we primarily used LUBsearch and Google Scholar, with additional guidance from our academic advisor, who directed us toward seminal papers. The literature review continued throughout the project to support our analysis and decision-making processes, allowing us to base our strategies and methodologies on well-established research findings.

The search strategy included terms such as "generative AI in corporate settings," "AI-driven documentation," and "automation of corporate processes". This approach helped us to focus on literature that discussed both the technological aspects of AI applications and their practical implications in business settings similar to the case company. We also explored methodologies relevant to the implementation of AI solutions, such as the CRISP-DM framework and design science in information technology, software engineering and prototyping.

To ensure a comprehensive review, we initially examined the abstract to determine whether the article was applicable to our project, focusing on its relevance to generative AI and documentation processes. If the abstract met our criteria, we proceeded to analyse the entire article, and categorised them based on a few categories; generative AI, corporate applications and AI-based automation techniques.

Regarding the selection process, we screened approximately 30 articles, assessing their relevance based on factors such as publication date and applicability to our study's context. After the broader review, we narrowed our focus to around 10 articles that directly addressed our key topics.

This ongoing literature review not only deepened our understanding of the subject but also ensured that our approach remained aligned with the latest developments and best practices in the field. It provided a critical backdrop against which we could evaluate the effectiveness of our proposed solution and refine our approaches based on proven research outcomes.

Semi-Structured Interviews

We conducted two semi-structured interviews with two knowledge specialists at the case company to gain a deeper understanding of the problem domain in the specific business context. The interview process was carried out in two stages. The first stage consisted of an exploratory interview particularly focusing on the documentation processes and the challenges faced by employees. Following the exploratory interview, a focused interview was organised with the same participants to clarify specific points and delve deeper into certain issues that had emerged. The detailed list of participants present during both interviews and their respective roles and experience levels can be found in Table 3.1.

Table 3.1: Participants of interviews. Column I denotes the identifier for each participant.

Current role at case company	Time worked at case company	Time worked within business domain	I
Knowledge Specialist	15 months	9 months	I1
Knowledge Specialist	4 years	6 months	I2

Exploratory Interview

The initial interview was an exploratory interview focused on gaining insights into the business operations and documentation practices within the case company. Prior to the interview, we conducted a brainstorming session to identify key areas that required further investigation, including underlying problems with the current documentation processes, the need for improvement, and the way employees interact with existing workflows. The interview was then conducted in a group setting via video conferencing, lasting approximately 45 minutes. The interviewees (I1, I2) are knowledge specialists with experience and responsibility in the processes of document generation and approval.

During the interview, detailed notes were taken by us to capture the key points discussed. Following the session, a summary was prepared, integrating the information collected from the notes. This summary was then organised according to the outline developed during our brainstorming session to ensure a coherent structure that aligned with the research objectives.

Focused Interview

Following the initial exploratory interview, we conducted a focused interview to explore specific details that had emerged during the earlier session. This focused approach aimed to clarify points of interest, resolve ambiguities, and gain additional insights to further guide our study. Before the session, we crafted a set of detailed questions (see Appendix A) designed to address the areas that required deeper investigation. The selection of these questions was driven by insights gained during the exploratory interview and informed by the gaps identified in our literature review.

The focused interview was also conducted via video conferencing but differed from the exploratory session in terms of structure and duration. This session was shorter, lasting approximately 30 minutes, with a more concentrated line of questioning. The interview was conducted with the same knowledge specialists (I1, I2). Notes were taken by the both of us during the focused interview and later compiled to a summary during a collaborative setting.

After the focused interview, we applied a thematic analysis to interpret the data collected. This allowed us to identify, analyse, and report patterns (themes) within the data. The first step in our analysis involved a careful reading and re-reading of the interview notes and summary to become familiar with the content. This immersion into the data helped us to notice patterns and themes emerging from the responses given by I1 and I2. We each independently identified crucial elements and codes from our analyses. Each code represented a concept or idea that appeared to capture something important about the data relevant to our research questions. Next, we grouped related codes into potential themes, continuously refining the specifics of each theme. This phase involved a lot of moving back and forth between the summary and our emerging themes, checking if the themes worked in relation to the coded extracts and the entire summary. Later, we discussed and chose the most fitting themes and codes, drawing on our separate insights.

The thematic analysis highlighted several key themes, including documentation practices, the impact of technology on documentation efficiency, and the challenges faced by employees in adapting to new documentation tools. These themes (see Table 4.1) were then discussed in relation to document generation and AI adoption in the workplace.

3.1.2 Data Understanding

Following our investigation of the business understanding, our next objective was to deepen our understanding of the specific data elements critical to the documentation, setting the stage for the design of a tailored solution. This data understanding phase was driven by document studies and focused on two main aspects: comprehending the distinctive tone of voice of the case company and identifying the typical structure of documents used within the organisation.

Tone of voice

The case company's brand has a distinctive tone of voice, which reflects the company's values, culture, and commitments. To understand the distinctive tone of voice, we conducted a document study centered on existing documents within the documentation database and reviewed official guidelines on the company's tone of voice. The analysis aimed to identify key linguistic and stylistic elements that define the case company's communication. This involved examining a selection of documents from the case company's documentation database to identify the common characteristics that represent the company's unique tone.

To select documents for this analysis, we worked with I1 to identify a representative sample of documents known for their consistent adherence to the case company's tone of voice. We reviewed a smaller subset of the database, approximately 10-20 documents, analysing them for recurring language patterns, vocabulary choices, and stylistic themes. The analysis was conducted collaboratively.

Document Structure

In our analysis of the document structure used within the case company, we conducted a document study aimed at uncovering the key characteristics that define the structure of the documents within the documentation database. The document study involved analysing a new subset of the database, approximately 40-50 documents, to identify common structural elements, such as titles, summaries, and content organisation. By focusing on these aspects, we were able to determine the typical patterns and layouts that characterise the company's documentation. This examination provided insights into how documents are structured, which the guidelines for our solution was based upon.

3.2 Define Objectives of a Solution

After completing the business and data understanding phases, we focused on defining the objectives of our AI-driven documentation solution. This step was informed by the insights gained from the business and data understanding. By synthesising this information, we were able to identify the key challenges and areas for improvement in the existing documentation processes.

The definition was done by brainstorming ideas with I1. This session was guided by the data we had gathered, focusing on the identified gaps and challenges in the current workflows. The goal was to develop objectives that were realistic and achievable within our available resources, considering factors such as time constraints, data accessibility, and technical capabilities.

3.3 Design and Development

In this step of our project, we developed a solution for generating documentation, based on objectives defined in the previous step. Our approach was structured into three components: Data Preparation, Model Development, and Implementation.

3.3.1 Data Preparation

We prepared data for training the GPT-3.5 model used in the solution through collecting, cleaning, and pre-processing a large amount of documents. It's important to note that this process was not repeated with GPT-4, as the fine-tuning feature was not available at the time we conducted this study. To address the distinct requirements of fine-tuning and Retrieval-augmented generation (RAG), we curated two separate datasets: one to capture the specific tone of voice and style of the case company's documentation, and another to cover a broader base of knowledge from the documentation database.

Fine-tuning

We collected a set of documents that are relevant to the domain of interest and of high quality to be suitable as training data for the fine-tuning process. To address this, we received examples from the I1 on relevant documents. These were then used as inspiration in order to manually select documents from the knowledge database to assemble the dataset.

Once the dataset was assembled, the next hurdle was cleaning and pre-processing of the data. Documents often contain a mix of textual content and multimedia elements. For the purpose of fine-tuning the model, it was necessary to strip away non-textual elements and standardise the format of the text. The next challenge involved segmenting the cleaned documents into smaller, coherent units that could be effectively used for training. These challenges was solved by implementing Python scripts that could automate the process of preparing the data.

For the fine-tuning process, a challenge was ensuring that the model could comprehend and generate text with an awareness of the overall context of the documents. This is particularly important in documentation, where details in one section can be crucial for interpreting information elsewhere. To address this, we developed a strategy where adjacent segments of text were provided to the model during training, allowing it to learn the flow of information and context transition within documents. Additionally, we experimented with including brief summaries or outlines of documents as part of the training data, thereby giving the model clues about the document's overall structure and content. The structure of text-prompts used to fine-tune the model for our first iteration is found in Listing 3.1.

Listing 3.1: The first draft of prompts given to the model for fine-tuning.

```
{
  "messages": [
    {
      "role": "system",
      "content": "<The context of the model was given here.
        This included context for being part of the case
        company’s documentation and instructions on how
        to provide the correct data. Restrictions were
        also provided in this context.>"
    },
    {
      "role": "user",
      "content": "<This part mimic the data given by the
        user to the model which is used for creating a
        suitable and correct document.>"
    },
    {
      "role": "assistant",
      "content": "<This part shows how the model is
        expected to respond to the users data given
        above.>"
    }
  ]
}
```

Retrieval-Augmented Generation

For the RAG dataset we vectorized the entirety of the knowledge database, approximately 50.000 documents by employing LangChain tools (see Section 2.1.4). This comprehensive vectorization process transformed all existing documents into vector representations that encapsulated their semantic essence, thus facilitating a nuanced understanding of the content. These vectors were then integrated into our purpose-built database, which was specially designed to support the RAG mechanism.

As previously mentioned, the dataset employed for the RAG approach was covering the entirety of documents in the documentation database. This holistic approach aimed to ensure that the model have access to a vast repository of information, thereby enabling it to pull from a wide array of topics and details to support the generation of content.

To further enhance the effectiveness of the RAG system, a search engine capability was integrated. This search engine functionality enables the model to access and incorporate current information available beyond the confines of the knowledge database. By enabling real-time retrieval of external data, this feature has the potential of increasing the model's ability to produce desired content. The integration of the search engine functionality thereby expanded the model's reach making it possible to handle a broader spectrum of queries and generating fitting responses.

3.3.2 Model Development

With a prepared dataset at hand, the project advanced into the model development phase. The processes of fine-tuning and prompt engineering was iterative, with continuous refinement based on testing feedback, aiming to strike an optimal balance between creativity and accuracy, ensuring the outputs were in harmony with the project's objectives.

Fine-Tuning

The fine-tuning process began by training the model on a curated dataset, which was divided randomly into training and validation sets at an 80/20 ratio. Adjusting the model's parameters was an iterative task, with each cycle intended to closer align the AI's outputs with the organisation's standards. The training and validation loss, which illustrate the model's learning progress and generalisation capabilities, are depicted in Figure 4.2 and Figure 4.3 respectively.

Through successive rounds of training and testing, the model gradually improved, demonstrating enhanced capability in producing documentation that met the predefined criteria.

Prompt Engineering

Parallel to fine-tuning, prompt engineering played an important role in the model development phase. This technique was utilised to guide the AI in generating content that not only adhered to the factual and stylistic requirements of the project but also aligned with the strategic objectives of the case company. Crafting effective prompts involved a deep understanding of how the AI interprets various instructions and the subsequent impact on the generated content. The iterative nature of prompt engineering allowed for the exploration of different prompt structures and contents, enabling the identification of optimal formats that consistently elicited the desired responses from the AI.

The process of prompt engineering began by choosing a prompt, followed by analysing different generated documents with this prompt. Each document was assessed on a scale of 1 to 10 for both language grade and factual grade. Subsequently, multiple different prompts were tested, and the resulting documents were evaluated using the same criteria. The feedback loop between prompt engineering and model training ensured a dynamic refinement of the AI's capabilities, resulting in a model that could produce high-quality, brand-aligned documentation efficiently.

3.3.3 Implementation

The implementation phase translated the theoretical model into a practical application capable of generating documentation. This involved integrating both the fine-tuned GPT-3.5-Turbo and the standard GPT-4 model into a user-friendly interface (see Appendix D.1), allowing for interaction and document generation for customer support. This phase also addresses operational considerations such as response time, scalability, and user feedback mechanisms to refine the model's performance further.

To achieve this, we selected the JavaScript framework Next.js as the foundation for our front-end interface. Known for its efficiency and flexibility, Next.js facilitated in the creation of a responsive, intuitive user interface that catered to the needs of users within the case company. This choice aimed to ensure that our application was not only accessible but also maintained a high standard of user experience, essential for encouraging adoption and engagement among the staff.

On the backend, we leveraged API routes powered by LangChain to connect the Next.js interface with our AI models. This setup utilised several key components from the LangChain library (see Section 2.1.4) to enhance functionality:

- *ChatOpenAI*: This component from the @langchain/openai package is used to interact with OpenAI's models.
- *SupabaseVectorStore*: Integrated from @langchain/community/vectorstores/supabase, this component is utilised to store and retrieve vector representations of documents in a Supabase database, facilitating data handling and retrieval.
- *PromptTemplate*: From @langchain/core/prompts, this component is used to structure and format the prompts sent to the AI model.
- *RunnableSequence*: This component from @langchain/core/runnables allows us to chain multiple operations in a sequence to process user inputs, interact with the model and parse outputs.
- *StringOutputParser & BytesOutputParser*: These parsers are utilised to handle and format the output from the AI model, transforming them into usable formats for further processing or response delivery. These components can be found in @langchain/core/output_parsers.

This setup was necessary for processing user queries, executing the prompt engineering techniques, and delivering the AI-generated documentation in an efficient manner. LangChain's capabilities allowed us to streamline the interaction between the user interface and the model.

3.4 Application Demonstration

Following the implementation phase, the project advanced to the demonstration of the application and assessment of its impact on the organisation. This stage was designed to evaluate the practical utility of the application, ensuring it effectively met the objectives outlined in the project's scope. To facilitate this, we engaged in a structured feedback loop with I1 and I2 (see Table 3.1), focusing on garnering insights from the end-users who would interact directly with the application.

3.4.1 Feedback Loop and Iterative Refinement

An important aspect of this phase was establishing a robust feedback loop with early users within the case company, particularly those involved in documentation and customer support roles. Their firsthand experiences gave valuable insights into the application's user interface and overall usability. This feedback highlighted several areas for improvement and provided constructive suggestions for enhancing the application's functionality. By addressing feedback related to response time, scalability, and the effectiveness of feedback mechanisms, we were able to iteratively refine the application. This process aimed to ensure that the application not only met the immediate needs of the organisation but also possessed the flexibility for future adaptation and expansion.

3.4.2 Demonstration Process

The demonstration process was structured to provide a comprehensive understanding of the application's performance in real-world scenarios. A series of demonstration sessions were organised, wherein the application was presented to a select group of end-users. These sessions aimed to showcase the application's capabilities, focusing on its user interface, the efficiency of documentation generation, and the alignment with the case company's tone of voice and content standards. The sessions served as a platform for live testing and evaluation, offering immediate and actionable feedback from potential end-users.

3.5 Application Evaluation

The application evaluation phase was designed to assess the suitability of our solution in supporting its intended objectives and to understand its impact on the case company's documentation processes. This comprehensive evaluation considered the objectives (see Section 3.2) and the overall user experience.

To conduct the evaluation, I1 and I2 (see Table 3.1) interacted with the application by inputting prompts into the application to observe how it generated the required content. This practical interaction aimed to assess the solution's performance in a real-world setting, focusing on the quality and relevance of the generated text. Users then provided structured feedback in a tabular format on their interactions with the application, assessing the following aspects:

- *Context*: An indication to determine if contextual information was appropriately included in the generated responses.

- *Mistakes*: The number of errors or inaccuracies observed in the output.

- *Factual Grade*: A numerical score (1-10) reflecting the factual accuracy of the generated content.
- *Language Grade*: A numerical score (1-10) evaluating the language quality, including tone, style, and coherence of the generated content.

Additional questions to capture qualitative insights or specific feedback were also included in a separate form. These can be found in Appendix B. The information gathered provided the basis for our application evaluation. By analysing user feedback and performance metrics, we were able to estimate the operational effectiveness of the application. This iterative feedback process was crucial for identifying areas for refinement and making improvements to enhance the application's usability and accuracy.

The evaluation process also included two user sessions with I1 and I2 where the application was demonstrated and its performance assessed in real-time. These sessions provided additional insights into the application's effectiveness and its potential impact on improving documentation workflows within the case company. The iterative refinements made in response to this feedback contributed to a more effective and user-friendly solution, ultimately validating the application's value in streamlining documentation processes.

The results of this evaluation are detailed in Section 4.4, providing a comprehensive overview of the application's performance and its impact on the case company's documentation practices.

Chapter 4

Results

By conducting interviews and integrating generative AI technologies, we have acquired insights into both the existing challenges and the potential enhancements within the company's documentation system. In this section, we organise our findings according to the structured analysis of business and data understanding, as well as the development and evaluation of AI-driven solutions.

4.1 Context Understanding

We have investigated the case company to gain an understanding of the business context and of the data related to their document management. This includes how the knowledge platform interacts with customer support and content management. The results are presented in this subsection.

4.1.1 Business Understanding

The findings from the thematic analysis based on the conducted focused interview (see Section 3.1.1) with I1 and 12 (see Table 3.1) are organized in Table 4.1, which showcases the distilled themes and corresponding codes of the analysis. The themes from the analysis are: *Documentation Workflow*, *Roles and Responsibilities*, *Manual Documentation Challenges*, *Tools and Software*, *Quality Control and Validation* and *Efficiency and Technology Suggestions*. These are denoted using *italics* throughout this section.

A recurring theme from the interviews was the *Documentation Workflow* (described in Section 1.3), where "...a coworker searches for an answer in the knowledge database but does not find an answer. Then they leave feedback saying 'I can't find an article (document) on this subject' or 'I can't find an answer to this question'" (I1). This feedback triggers a sequence where specialists create or edit content based on the feedback.

Table 4.1: Themes and codes from the analysis of the focused interview. The themes are denoted in bold.

Documentation Workflow	
• Identify the need for content	• Submit feedback
• Create or update content	• Dynamic evolution
Roles and Responsibilities	
• Generalists identify and report gaps	• Specialists/Publishers create and update content
• Knowledge specialists manage the process end-to-end	• Customer support voice concerns
Manual Documentation Challenges	
• Time-consuming to identify and guide specialists	• Difficult to complete tasks promptly
• Grammatical errors	• Lack of adherence to tone of voice
• Resistance to new tools	• Resistance to new documentation types
• Bottlenecks because of manual intervention	• Process latency
• Human error creates inconsistencies	• Metadata application across markets
• Lack of adherence to guidelines and templates	
Tools and Software	
• Knowledge platform, Microsoft Office, SharePoint, Adobe tools, and ChatGPT	• Scattered nature
• Grammatical aid	• Content comparison
• Cross-checking	• Data sharing
Quality Control and Validation	
• Feedback-driven updates	• Four-eye principle
Efficiency and Technology Suggestions	
• Centralised documentation tools	• Unified platform
• AI integration	• Aiding individuals with disabilities

Regarding *Roles and Responsibilities*, it was noted that customer support agents, who have direct contact with customers, are often the first to voice concerns. Following these initial alerts, generalists identify specific areas that require updates or new content. Specialists/publishers are then tasked with the actual content creation, ensuring the process is complete and accurate. Knowledge specialists play a crucial role in overseeing this entire process, ensuring that all parts of the organisation effectively use tools and resources.

Challenges in Documentation Management were also identified, with these findings detailed in section 4.2, as this theme directly intersects with one of our research questions.

When discussing *Tools and Software*, one of the interviewees mentioned that the current tools in use include the Knowledge platform, Microsoft Office, SharePoint, Adobe tools, and ChatGPT. He elaborated on the challenges associated with multiple SharePoint instances by noting, "The problem with having multiple sharepoints is for example the access part that is really hard to keep updated and sharepoint doesn't have all the great functions the knowledge platform has" (I2). Limitations of the knowledge platform were specifically noted, particularly its inadequate support in handling grammatical errors and ensuring consistency in the presentation of information. It was also observed that although ChatGPT is a tool suited for revising nearly completed drafts, its effectiveness is limited by the models' lack of sufficient context and tone of voice.

In terms of *Quality Control and Validation*, the feedback-driven update process is central. One interviewee explained, "Generalists provide feedback if they find something wrong, outdated, confusing et cetera and the specialists picks it up and fixes it. We have a four eye-principle before publishing the content which works well" (I1). Despite these measures, consistency remains a challenge, underscoring a need for more robust mechanisms.

When discussing *Efficiency and Technology Suggestions*, one of the interviewees expressed optimism about AI, stating, "I personally believe that AI could solve issues we are facing with consistency and adherence to the company's tone of voice in the process of content creation" (I1). This reflects a broader desire for solutions that could streamline the documentation process, reduce errors, and improve inclusivity.

4.1.2 Data Understanding

We performed document studies to deepen our comprehension of the distinctive tone of voice of the case company and identifying the typical structure of documents used within the organisation. The results of our studies are presented in the subsequent sections.

Tone of Voice

We identified five core attributes that characterise the tone of voice of the case company: *straightforward*, *friendly*, *practical*, *diverse* and *responsible*.

- *Straightforward*: The case company uses language that is straightforward and designed for easy understanding.
- *Friendly*: The tone is consistently friendly, which makes the company feel approachable and warm. This is evident through the use of conversational language the use of language that involves the customer directly, using pronouns such as "we" and "you".
- *Practical*: The language often focuses on solving problems or providing practical advice.
- *Diverse*: The tone is inclusive, aiming to speak to a wide, global audience while respecting cultural differences. In many cases we found that the documents included phrases such as "...embracing all cultures and lifestyles"
- *Responsible*: The communication also incorporates elements of corporate responsibility such as sustainability.

Document Structure

We concluded that the documents are categorised in two main document types, external and internal. The two different types have different structures and content, but share the shape of the additional information provided in some documents.

Internal Documents

Internal documents focus on information that is more inclined towards employees and the information is not intended to be directly shared with customers. The internal documents generally do not contain references to external links, external information intended for customers or a wider source reference as these attributes are more applicable to external documents. Another characteristic of the internal documents is that they rarely tend to answer a specific question, instead opting to provide general knowledge of an area. A generalisation of the document structure of the internal documents can be seen in Table 4.2.

Table 4.2: Generalisation of the internal document structure.

Section	Content
Title	Title formed as summary of the document.
Summary	Summarised information in regards to the content of the document. Important information such as deadlines, requirements and more are written here.
How?	This part of the document is generally intended to explain who this document affects and how. This section could also include helpful links that are associated to the content of the document.
Additional information	Other information not covered in the "How?" section. It often includes a set of commonly asked questions regarding the content of the document and the answer to theses questions.

External Documents

The external documents are instead inclined on solving a specific question or use-case. The title is set up to pose the potential question customers ask, and the content contain information to make customer support able to answer the question sufficiently. It is a frequent practice for documents to include web links that customer support can share with customers, offering them an opportunity to delve deeper into the subject matter.

The language in the summarised part of the document tend to follow the tone of voice of the case company, which enables customer support to directly use the text and pass it on to the customer. External documents occasionally feature sections of internal content. These segments are designed to inform customer support agents about details that should not be disclosed to customers.

A generalisation of the document structure of the external documents can be seen in Table 4.3.

Table 4.3: Generalisation of the external document structure.

Section	Content
Title	Title formed as a question.
Summary	Summarised answer or information in regards to the title. This answer is meant to be directed towards the custom, with relevant tone of voice and suitable material.
Links	Potential links that could be useful for customer support or the customer.
Internal content	This optional part can contain information in a more direct way not suitable for customers. It could also include examples of different ways of answering the question. It is not structured in the company's tone of voice.

Metadata

Both the internal and external documents contain optional but preferred metadata based on the following fields: *Categories/Supervisor/Validity*, *Feedbacks*, *Keywords*, *Document information/History*, *Search phrases*, *Versions*, *Original of market specific variant*, and *Referencing Documents*. A description for each field can be found in Table 4.4.

Table 4.4: The optional metadata connected to each document.

Section	Content
Categories/Supervisor/Validity	This section contains information regarding how the documents have been categorised in the form of root navigation. It also shows information regarding who or which market is responsible for guaranteeing its validity. It could also include the timespan in which this information is valid.
Feedbacks	This section show the comments a co-worker can make on a document. This can include requests for changes, clarifications or other form of feedback.
Keywords	This includes specific words that the author can add in order to be able to find the document more effectively through the search function, as well as being able to filter and categorise documents.
Document information	Displays a variety of document data, such as type, status, publishing information, version no, author and more. It also displays the history log of who created, edited and resubmitted the document.
Search phrases	This includes specific phrases that the author can add in order to be able to find the document more effectively through the search function.
Versions	Displays links to the different published versions as well as the author and date of the different versions.
Original of market variant	This displays a link and some general information regarding the reference that the document was based upon.
Referencing Documents	This section contains information regarding potential references to other documents within this document. It also fetches and display the data of the document referenced, such as created by, doc.ID and published date.

4.2 Manual Documentation Challenges (RQ1)

The thematic analysis conducted from the semi-structured interviews provided insights into the challenges faced in the manual documentation processes at the case company. This analysis highlighted several areas where inefficiencies, delays, and discrepancies impact the overall effectiveness and reliability of the documentation system.

The following sections detail the primary challenges identified through our interviews, which include scalability and efficiency, time-consuming processes, inaccuracies and inconsistencies in the documentation, and resistance to adopting new technologies.

4.2.1 Scalability and Efficiency

A bottleneck within the case company's documentation processes is the reliance on manual efforts by customer support agents, specialists and publishers. This is noted by one interviewee: "It is time-consuming to find the right specialist and guide them through content creation tools and processes" (I1). This dependence not only limits the system's scalability but also raises concerns about its overall efficiency. The process's manual nature can delay responses to customer inquiries, impacting both customer satisfaction and the quality of service provided. "It takes a lot of time upskilling and reviewing the content that the publishers create" (I2) one interviewee noted, emphasising how these time-intensive manual processes further aggravating the challenges.

4.2.2 Time-Consuming Processes

The process from recognising the need for new content to its eventual publication involves a complex, multi-staged procedure that can be time-consuming. According to one specialist, "From our point of view, it is time-consuming to find the right specialist and guide them through content creation tools and processes. It is also time-consuming to push publishers to accept certain tasks and finish them as soon as possible" (I1). This coordination complicates the ability to swiftly address emerging customer requirements.

4.2.3 Inaccuracies and Inconsistencies

The reliance on human-authored content within the case company's documentation framework introduces inherent risks of inaccuracies and inconsistencies. This risk is particularly pronounced when contributions come from a diverse group of individuals across various regions, each with their own unique perspectives, writing styles and experience in writing documents. "Grammatical errors are common, but we also see that our coworkers generally lack understanding of the tone of voice in both internal and external content creation" (I1) as noted in our findings. These discrepancies can extend beyond textual semantics to include metadata management, where personal biases in categorisation, tagging, and other descriptors may arise. Such inconsistencies not only confuse customer support agents but also customers themselves, hindering their ability to navigate and utilise the documentation effectively.

4.2.4 Resistance to New Technologies

There is a noticeable resistance among the specialists towards adopting new technologies aimed at streamlining the documentation process. This reluctance poses challenges to the organisation's efforts to improve documentation handling efficiency and quality. "There is so much potential in AI..." (I2) remarked one interviewee, while the other interviewee also explains that "...it is difficult to get them (specialists) to test new tools [...] and motivate them to upskill themselves".

4.3 Generative AI for Documentation (RQ2)

We built an application called Peyhal Platform to enable the use of generative AI in documentation creation at the case company. The approach and development of the application are detailed in Table 3.1 (see Section 3). The system context for this application can be seen in Figure 4.1 and a more detailed technical container context for the application can be found in Appendix D, Figure D.2.

Several design decisions were made to navigate the complexities of AI-based content creation. These decisions culminated in the adoption of a hybrid approach that combines fine-tuning, RAG, and prompt engineering techniques (see Section 2.1.2). By opting for this mixed approach, we aimed to create a solution that combines the personalised accuracy of fine-tuned models with the contextual richness and real-time updating capabilities of RAG. This design decision reflects a strategic choice to build a system that is not only capable of producing high-quality documentation but also adaptable, scalable, and capable of evolving in tandem with the organisation's changing needs.

4.3.1 Fine-Tuning

The decision to incorporate fine-tuning as a core component of one of the models in our approach was motivated by the necessity to tailor the AI model to the specific linguistic style, terminology, and content preferences of the organisation. By adjusting the model parameters based on a curated dataset of existing documentation, we ensured that the generated content would not only be relevant and informative but also reflect the company's tone of voice and adherence to its guidelines. This process of fine-tuning allows the model to produce outputs that are more aligned with the organisation's standards than those generated by a general-purpose AI model.

The result in terms of training loss and validation loss from the fine-tuning can be found in Figure 4.2 and Figure 4.3 respectively. The trend line depicted in the figure indicates a slight decrease in validation loss over iterations of fine-tuning, suggesting improvements in model performance on unseen data. Furthermore, the training loss exhibits a slightly more pronounced decrease compared to the validation loss. This behaviour indicates that the model is effectively learning and adapting from the training data, optimising its parameters to minimise the error on the training set while still generalising well to the validation set.

4.3.2 Retrieval-Augmented Generation

Complementing the fine-tuning, we integrated RAG elements into our solution, utilising both a vector database constructed from previously written documents in the knowledge database and a search engine capability. This dual RAG approach enables the AI to dynamically retrieve and utilise existing documentation as a reference point during the content generation process. The vector database provides a robust mechanism for identifying and leveraging relevant documents, ensuring that the generated content is consistent with previously established knowledge and information.

The inclusion of search engine functionality allows the model to pull in the most up-to-date information from a broader set of external sources, enhancing the accuracy and comprehensiveness of the content.

4.3.3 Prompt Engineering

An additional design decision in our project was the focus on prompt engineering as a method to optimise the interactions with our AI model, particularly in the context of generating documentation. This technique became crucial for refining how the model interprets and acts on the input data, especially when aiming for specific styles, tones or formats of documentation. By designing prompts that align with our objectives, we were able to more precisely control the content generation process, ensuring that the outputs not only met the informational needs but also adhered to the linguistic and stylistic requirements of the case company. This iterative approach of designing new prompts improved the accuracy and relevance of the generated documentation. The outcomes from the prompt engineering iterations (see Section 3.3.2) are summarised in Table 4.5, demonstrating the impact of tailored interactions in achieving desired AI-generated content. The prompts used in these iterations can be found in Appendix C. Because of the generally better result, Prompt 5 was chosen for the quality and accuracy evaluation of the application.

Table 4.5: Evaluation of different prompts ranked from 1 to 10 on language and factual accuracy. The average result was derived from a total of 90 generated documents.

Prompt Version	Language Grade	Factual Grade
Prompt 1	3.3	3.4
Prompt 2	4.7	5.2
Prompt 3	6.1	5.2
Prompt 4	6.7	7.4
Prompt 5	7.1	7.2

4.4 Quality and Accuracy Evaluation (RQ3)

The application evaluation (see Section 3.5), was based on two models (see Table 4.6) to provide a deeper understanding of the impact of contextual information on model performance.

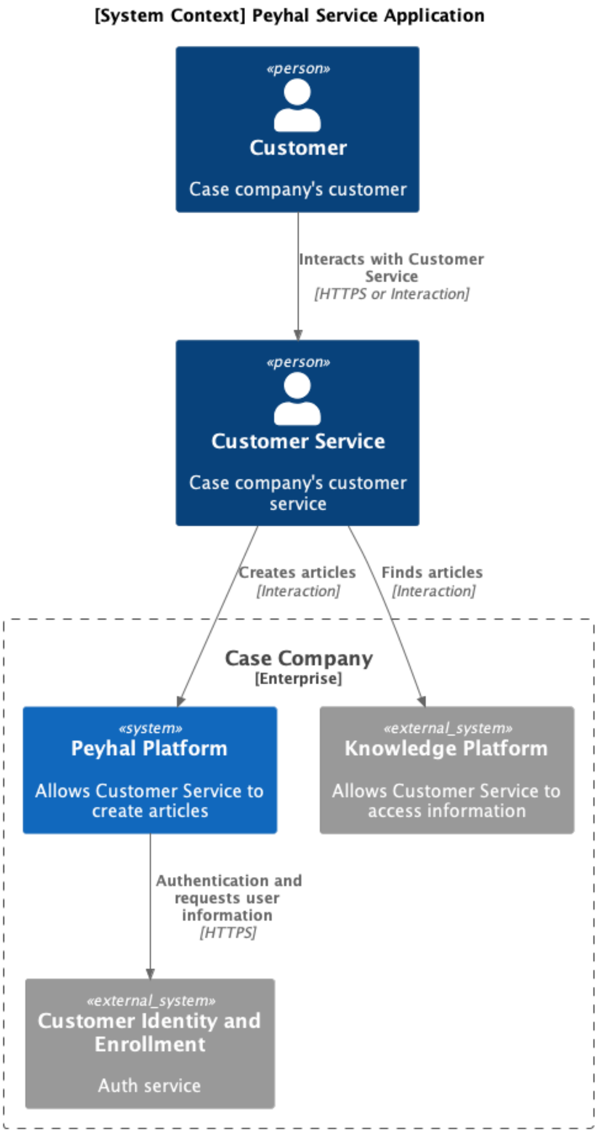

[System Context] Peyhal Service Application
«person»
Customer
Case company's customer
Interacts with Customer Service
[HTTPS or Interaction]
«person»
Customer Service
Case company's customer service
Creates articles
[Interaction]
Finds articles
[Interaction]
Case Company
[Enterprise]
«system»
Peyhal Platform
Allows Customer Service to create articles
«external_system»
Knowledge Platform
Allows Customer Service to access information
Authentication and requests user information
[HTTPS]
«external_system»
Customer Identity and Enrollment
Auth service

Figure 4.1: An overview of the integration of Peyhal Platform within context of the case company.

It focused on minimising mistakes, ensuring factual accuracy, and enhancing linguistic quality. The evaluation consisted of a combination of a performance measurement of the form assessment (see Section 3.5) and by letting two case company employees use the implemented application to provide feedback.

Table 4.6: The models evaluated for this thesis indicated by their version, snapshot and whether they underwent fine-tuning.

GPT-version	Model snapshot	Fine-Tuned	Identifier
GPT-3.5-Turbo	1106	Yes	M1
GPT-4	0125	No	M2

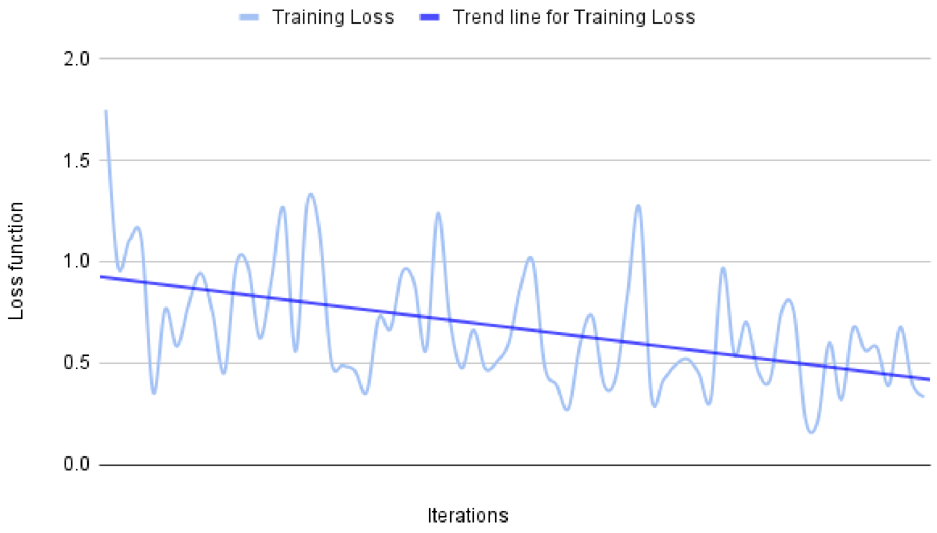


Figure 4.2: Training loss for the Fine-tuning process.

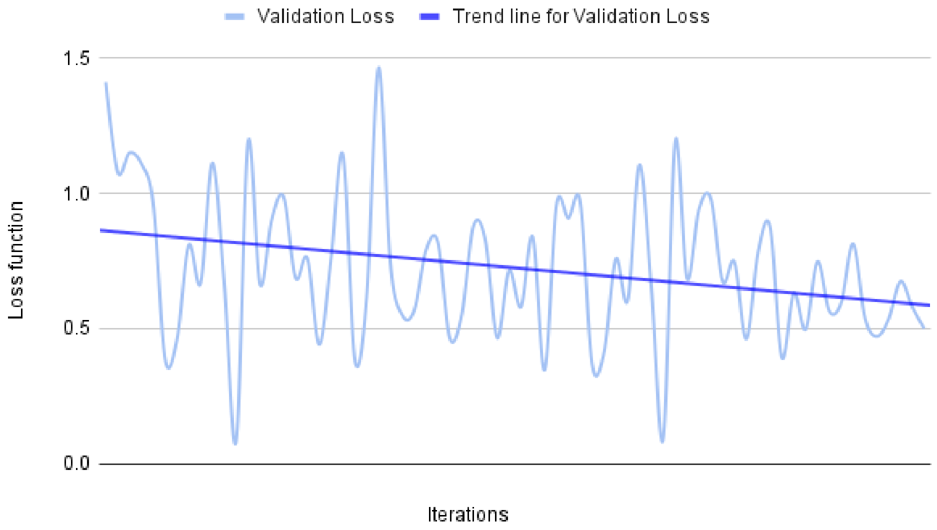


Figure 4.3: Validation loss for the Fine-tuning process.

4.4.1 Performance Measurements

The results from evaluating two models, M1 and M2, with or without context, are detailed in terms of their average number of mistakes and grading for factual and language quality from 1 to 10. The results are summarised in Figure 4.4 and Figure 4.5. Furthermore, Table 4.7 presents the same evaluation, incorporating whether context was used in generating the documentation. These results show that model M2 perform better than model M1 when it comes to minimising mistakes in both contexts. Specifically, M2 had fewer mistakes on average (2.5) in the absence of additional context compared to M1 (3.5). This pattern persisted with context, although the gap expanded slightly (M2 at 2.6 vs. M1 at 4.4). Regarding factual accuracy, M2 consistently outperformed M1. M2 achieved a higher average factual grade in both contexts, scoring 7.8 without context and 7.9 with context. In contrast, M1's factual accuracy suffered notably when context was provided, dropping from 6.8 to 5.1, suggesting potential challenges in integrating contextual information effectively. The language proficiency grades revealed a different trend. M1's language grade improved from 6.0 to 8.0 when context was provided, suggesting that the inclusion of context might help M1 generate more linguistically refined outputs. On the other hand, M2 showed less pronounced improvement but maintained consistently higher language grades than M1 without context (6.8 vs. 6.0). With context, M2 improved to a grade of 7.7. These findings suggest that while M2 is generally more adept at reducing mistakes and maintaining factual accuracy, M1 may benefit more in linguistic performance from the inclusion of additional context.

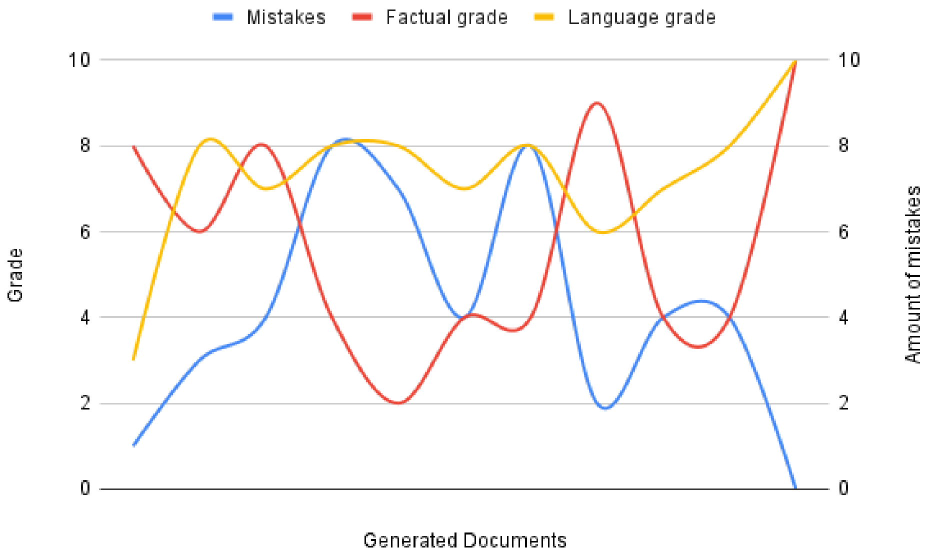


Figure 4.4: Evaluation result for the M1-model.

Table 4.7: Evaluation of the models M1 and M2 with and without context. The average result was derived from a total of 50 generated documents.

Model	Context	Mistakes	Factual Grade	Language Grade
M1	No	3.5	6.8	6
M1	Yes	4.4	5.1	8
M2	No	2.5	7.8	6.8
M2	Yes	2.6	7.9	7.7

4.4.2 Additional Feedback

Feedback insights from demonstration process with I1 and I2 who interacted with both models, M1 and M2, provide additional context to the performance differences and potential applications of these models. The feedback from the participants underscores the potential

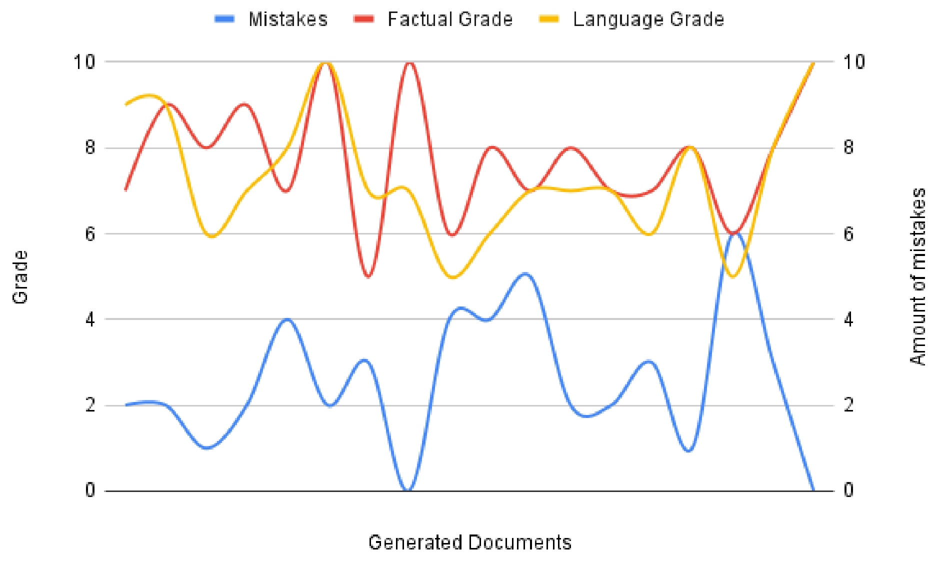


Figure 4.5: Evaluation result for the M2-model.

of both models to enhance content creation processes within the knowledge platform, improving both the efficiency and quality of outputs. While M2 was highlighted for its more reliable factual accuracy, M1 was commended for its language capabilities, as can be seen from one participant's response "...the main differences are that the M2 seems to be better when it comes to factuality and M1 better at language" (I1). This statement was further complemented by the other participant by stating: "...there was more errors in articles (documents) created by M1" (I2).

M1 displayed potential to enhance the quality of documents in the knowledge database, though it sometimes struggled to follow detailed instructions, indicating a need for improvements in its ability to parse and respond to user input. The following was stated by participant I2, "...it (M1) has a tendency to answer 'around' the question". Its performance was irregular when context was provided, suggesting that its integration of contextual information could be more effectively optimised. M2 was perceived as more consistently impressive, particularly noted for its stronger handling of context. One participant (I1) described M2 as "...I think this model is very promising". I1 were also impressed by its detailed explanation of some concepts, showcasing the model's robust handling of specific queries. The general consensus was that M2 performed better with the addition of context. When asked if the model perform better given context, one participant responded: "...generally I would say yes" (I1).

Both models were considered an improvement over the existing average document in the knowledge database, particularly in terms of tone and spelling, aligning closely with the company tone of voice as this can be understood from one participant commenting on both models, "...tone of voice and spelling is much better than the average document currently existing in the knowledge database" (I1). This opinion was partially shared by participant I2 who stated, "Yes, the grammar errors is mostly gone using the AI tool". There were some minor concerns about M1's tendency to hallucinate, though these instances were not frequent.

Chapter 5

Discussion

In this chapter, we examine and interpret the results from our study on the documentation processes at the case company, assessing the effectiveness of the proposed solutions and exploring potential areas for future advancements. Our review highlights where the intended objectives were achieved and identifies opportunities for further development and improvement.

5.1 Manual Documentation Challenges (RQ1)

The thematic analysis from the interviews at the case company provides a depiction of the challenges faced in manual documentation processes: *scalability and efficiency*, *time consumption*, *inaccuracies and inconsistencies*, and *resistance to technology adoption*. Discussing these results further, we can deepen our understanding of implications and the potential strategies for improvement, as well as the risks associated with inaction.

Scalability and efficiency challenges in manual documentation processes strain resources and jeopardise operational agility and responsiveness. Embracing automation and digital tools becomes crucial for organisations looking to overcome these challenges.

The time-consuming nature of manual documentation processes reflects broader inefficiencies in project management and communication protocols. Streamlining these processes through technological interventions can lead to gains in operational efficiency and agility.

The prevalence of inaccuracies and inconsistencies in manual documentation underscores the inherent limitations of relying solely on human inputs. The risk of errors in documentation poses a challenge to maintaining trust and reliability with customers. Integrating advanced quality control technologies, such as AI-driven analytics and natural language processing tools, can mitigate these risks by ensuring accuracy and consistency across documents. By embracing such technologies, organisations can not only enhance the quality of their documentation but also bolster their reputation for reliability and precision.

The resistance towards adopting new technologies represents a cultural barrier that transcends individual organisations. Based on the interviews, we argue that the resistance is based on a fear of the unknown, concerns over job security, and the potential learning curve associated with new systems. There may also be a disconnect between leadership and staff regarding the perceived benefits of technology, where leaders see potential gains in efficiency and innovation, while staff may view these changes as disruptive and unnecessary. Such a gap between perceived and actual benefits can hinder technological adoption, stalling progress and innovation within the organisation. Adding to the complexity of this issue, there's an argument to be made that new solutions often meet a higher requested standard than traditional ones. This point invites a broader discussion on the actual versus perceived impact of AI.

5.2 Generative AI for Documentation (RQ2)

The integration of generative AI into documentation processes represents an advancement in how companies manage and produce their corporate content. Our application built in cooperation with the case company demonstrates not only the technical feasibility but also the strategic advantages of adopting such technologies.

The adoption of a hybrid AI approach, which combines fine-tuning, Retrieval-Augmented Generation (RAG), and prompt engineering, suggests a robust method for achieving tailored and contextually appropriate documentation. This strategy implies that the generated content is not only accurate and relevant but also finely aligned with the specific linguistic and stylistic nuances of the organisation. This tailored approach to AI documentation generates several potential benefits, including increased efficiency in content creation and the ability to maintain a consistent company specific voice across all documents.

The implementation of fine-tuning specifically addresses the necessity for customisation in corporate environments where brand consistency is crucial. By training the AI on a dataset that reflects the organisation's existing documentation, the AI model adapts to replicate the organisation's unique style and terminology. This suggests that generative AI can become an integral part of an organisation's content strategy, potentially reducing the need for extensive manual revisions and ensuring that all documentation is on-brand and aligned with corporate standards.

The use of RAG for content generation emphasises the importance of context and relevance in AI-generated documentation. By utilising a vector database of documents along with real-time search engine capabilities, the AI solution can pull from a vast amount of both internal and external information, ensuring that documents are not only consistent with previous content but also current with the latest information. This aspect of the AI application suggests a reduction in the resources typically required to keep documentation up-to-date and relevant, thus offering a compelling case for its adoption in environments where information is continually evolving.

The emphasis on prompt engineering within our application demonstrates a level of control over the AI's output, enabling precise customisation of content to fulfil specific requirements. This approach can enhance the relevance and utility of AI-generated documentation, making it possible to produce content that meets specific user requirements without extensive human intervention. As AI technologies advance, the role of human oversight could shift from content creation to more strategic activities such as prompt optimisation and result evaluation, highlighting a shift in job roles and skills required within the documentation field.

5.2.1 Implementation Improvements

As we consider the integration of generative AI into documentation processes and reflect on the implemented application, there are several potential improvements that could enhance the effectiveness and efficiency of this AI-driven approach. These enhancements aim to refine the system's performance and extend its capabilities to better meet the organisational needs.

Reinforcement Learning

One improvement involves the implementation of reinforcement learning techniques. By incorporating reinforcement learning, the AI system can continuously learn and improve from its interactions and outputs based on feedback. This method would enable the AI to adjust its content generation strategies over time, optimising for better alignment with user satisfaction and organisational objectives. This iterative learning process not only refines the AI's accuracy and relevance but also helps in adapting to changing documentation requirements and preferences, ensuring the system remains dynamic and effective.

Expanding the Vector Database

Another area for improvement is the expansion of the vector database to include more comprehensive information from various data sources within the company. By broadening the scope of the database, the AI can access a wider array of reference materials, which enhances its ability to generate content that is both contextually rich and highly relevant. Incorporating documents, reports, and data from different departments could provide a more holistic view and understanding, enabling the AI to produce documentation that better reflects the full spectrum of the company's operations and knowledge base.

Search Engine Algorithm

Improving the efficiency of the search engine algorithm used by the AI to find relevant content is a potential area of improvement. By refining this algorithm, the system can more quickly and accurately identify the most appropriate existing content to use as references. This adjustment would reduce the time spent retrieving information and increase the accuracy of the content generation process. Enhancements might include better keyword recognition, understanding of context, and integration of semantic search techniques, which together would streamline the content creation workflow.

5.3 Quality and Accuracy Evaluation (RQ3)

The results from the evaluation offer insights into the quality and accuracy of the two models, M1 and M2 (see Table 4.6), in generating content within the knowledge framework. Our evaluation aimed to give a broader understanding of how contextual information influences model performance, particularly focusing on minimising errors, ensuring factual accuracy, and enhancing linguistic quality.

Our analysis indicates that the presence of contextual information impacts the quality and accuracy of the outputs generated by AI models. Model M2, which was not fine-tuned, demonstrated improved performance in maintaining factual accuracy and minimising mistakes regardless of the context provided. This suggests that more recent or advanced iterations of generative AI models may have inherent capabilities in handling factual information more reliably. On the other hand, Model M1, which was fine-tuned, showed an improvement in linguistic quality when context was provided. This indicates that fine-tuning on specific contexts can enhance a model's ability to produce linguistically refined outputs, which is important for applications requiring a high level of language proficiency, such as content creation in nuanced fields.

The performance discrepancy between M1 and M2, particularly with the integration of contextual information, highlights an essential consideration for the deployment of AI in producing documentation: the trade-off between factual accuracy and linguistic enhancement. While M2 consistently showed fewer errors and higher factual grades, M1's performance in language quality suggests that there might be benefits to using contextually fine-tuned models for tasks where linguistic style and nuance are prioritised.

From the additional feedback provided by the participants interacting with both models, it is apparent that while M2 was preferred for tasks requiring strict factual accuracy, M1 was favoured for its enhanced language capabilities, which may be beneficial in fields like creative writing or content tailored to specific brand voices.

The varied performance of the models with and without context underlines the potential of contextual fine-tuning in improving the quality of AI-generated content. As AI models become more integrated into different corporate sectors, understanding the effects of contextual information will be central in leveraging these technologies effectively. It is also notable that while M2 performed better overall, the integration of contextual information led to improvements in both models, suggesting that context is a viable enhancer of performance.

The findings of the assessment highlight the need for ongoing evaluation and refinement of AI models, as well as the promising prospect of AI's capability to produce accurate and high-quality documentation. Though there is room for enhancements and adjustments, the overall reception was optimistic.

5.4 Threats to validity

Addressing and discussing the threats to validity inherent in any study is not just a matter of transparency, but also essential for increasing the credibility of the research. By acknowledging the potential sources of bias, error, or limitations, we hope to demonstrate and enhance the trustworthiness of the findings in this thesis.

The large and extensive data utilised for the RAG and fine-tuning processes presents a threat to validity. The lack of validation mechanisms for ensuring the quality and validity of this dataset raises concerns regarding its reliability and accuracy in reflecting real-world scenarios accurately. There is a possibility that some of the data is inaccurate, out of date or simply not up to standard. AI-models can be described as a reflection of the data they are based upon, which puts the result of our thesis at a validity risk.

Our reliance on only two participants for both interviews aimed at evaluation and understanding business needs, as well as the assessment of our AI application, introduces a potential threat to the internal validity of our findings. With the narrow participant pool, there is a risk that our results may not fully capture the broader perspectives, opinions and experiences across the entire case company.

There is also a possibility of bias in our thematic analysis, as we conducted it ourselves. The subjective interpretation of themes and patterns within the data could be influenced by preconceived notions or expectations, leading to potential distortions in the findings and conclusions drawn from the analysis. To minimise the risk of bias, we each independently pinpointed key points and developed themes and codes from our analyses. Subsequently, we came together to discuss and select the most appropriate themes and codes based on our individual insights.

The reliance on human evaluation as the primary basis for deriving conclusions introduces a threat to the internal validity of our study. Human evaluators may exhibit variability in their assessments, leading to inconsistencies and potential inaccuracies in the interpretation of results. The subjective nature of human judgement may also introduce bias into the evaluation process, undermining the objectivity and reliability of our findings. Since the evaluators were aware of the model they were using, there is a possibility of bias in their interpretation of the results. This knowledge could influence their expectations and perceptions, potentially leading them to consciously or unconsciously favour certain outcomes.

Given that the research has been conducted within a particular organisational context, the results may not be directly applicable to other entities. However, general results of our study suggest that the findings could indeed have broader relevance. Cases that share similar issues with manual documentation processes could see the application as a potential solution or inspiration. The evaluation results based on the different approaches such as fine-tuning, RAG, prompt engineering, context and more could provide general insights into any case looking to adopt an AI-solution.

Chapter 6
Conclusion

Our thesis aimed to explore the impact of integrating generative AI into the documentation processes at a case company, examining both the effectiveness of this integration and the challenges encountered. We conducted an investigation into the current documentation practices, developed an AI-driven solution, and analysed the results to understand how AI can enhance content creation and management.

In our investigation of manual documentation processes within a corporate setting (RQ1), several challenges were identified. These included scalability issues, inefficiencies, resistance to new technologies and inconsistencies in the company's documentation practices. These issues underscore the need for a more efficient way of managing documentation. By exploring how documentation can be automatically created using generative AI (RQ2), we developed and implemented an AI solution that automates part of the documentation process. This solution uses techniques such as fine-tuning, Retrieval-Augmented Generation, and prompt engineering to ensure the generated documents are not only accurate but also align with the company's stylistic and informational standards. The subsequent evaluation of the documentation produced by the AI solution (RQ3) showed that it is possible to generate documentation that meet the company's requirements for compliance and style. The evaluations also showed an improvement in the documents consistency and factual accuracy depending on the techniques and models used.

Insights gained from the case company's feedback indicate a positive outlook towards the adoption of AI technologies for future operations. The company has expressed interest in expanding the AI application beyond documentation to other areas of business where automated data processing and management can increase efficiency. A longer-term deployment of the application would provide deeper insights into its performance over time, particularly in how well it adapts to evolving business needs and documentation standards. Such extended deployment should also investigate the AI's impact on broader organisational processes and include continuous user feedback to refine AI outputs.

We opted not to include multimodal generation in our work, which would involve integrating images, graphs and other media to potentially enhance the quality and usefulness of the documentation. Although this presents an interesting area for exploration, it poses challenges in achieving seamless and effective incorporation. With newer models like M2 demonstrating generally superior capabilities, there is also a need to assess the diminishing returns of fine-tuning earlier models like M1. This analysis could provide valuable insights into the balance between adopting new technologies and optimising existing ones.

Our findings contribute to a broader understanding of the potentials and limitations of applying AI in business processes, particularly in the realm of corporate documentation. This research can serve as a reference for similar initiatives, helping other organisations navigate the complexities of digital transformation in their documentation practices.

References

[1] Hervé Abdi, Dominique Valentin, and Betty Edelman. *Neural networks*. Sage, 1999.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Ana Azevedo and Manuel Filipe Santos. Kdd, semma and crisp-dm: a parallel overview. *IADS-DM*, 2008.

[4] Angels Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Roberto de M Estevão Filho, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O Nunes, et al. Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. *arXiv e-prints*, pages arXiv–2401, 2024.

[5] Ertuğrul Bayraktar, Cihat Bora Yigit, and Pinar Boyraz. Tailoring the ai for robotics: Fine-tuning predefined deep convolutional neural network model for a narrower class of objects. In *Proceedings of the 2017 International Conference on Mechatronics Systems and Control Engineering*, pages 10–14, 2017.

[6] Yoshua Bengio and Yann LeCun. Scaling learning algorithms toward ai. 2007.

[7] Noémi Bontridder and Yves Poullet. The role of artificial intelligence in disinformation. *Data &#38; Policy*, 3:e32, 2021.

[8] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, jan 2024.

[9] K. R. Chowdhary. *Natural Language Processing*, pages 603–649. Springer India, New Delhi, 2020.

[10] Michael Chui, James Manyika, and Mehdi Miremadi. What ai can and can't do (yet) for your business. *McKinsey Quarterly*, 1(97-108):1, 2018.

[11] Jari Dahmen, M Enes Kayaalp, Matthieu Ollivier, Ayoosh Pareek, Michael T Hirschmann, Jon Karlsson, and Philipp W Winkler. Artificial intelligence bot chatgpt in medical research: the potential game changer as a double-edged sword. *Knee Surgery, Sports Traumatology, Arthroscopy*, 31(4):1187–1189, 2023.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] X Du-Harpur, FM Watt, NM Luscombe, and MD Lynch. What is ai? applications of artificial intelligence to dermatology. *British Journal of Dermatology*, 183(3):423–430, 2020.

[14] Issam El Naqa and Martin J. Murphy. *What Is Machine Learning?*, pages 3–11. Springer International Publishing, Cham, 2015.

[15] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 201–208. JMLR Workshop and Conference Proceedings, 2010.

[16] Damien Ernst and Arthur Louette. Introduction to reinforcement learning. 2024.

[17] Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. Generative ai. *Business & Information Systems Engineering*, 66(1):111–126, 2024.

[18] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

[19] Kevin Gurney. *An introduction to neural networks.* CRC press, 2018.

[20] Abid Haleem, Mohd Javaid, Mohd Asim Qadri, Ravi Pratap Singh, and Rajiv Suman. Artificial intelligence (ai) applications for marketing: A literature-based study. *International Journal of Intelligent Networks*, 2022.

[21] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International conference on machine learning*, pages 2712–2721. PMLR, 2019.

[22] Yayan Heryanto and Agung Triayudi. Evaluating text quality of gpt engine davinci-003 and gpt engine davinci generation using bleu score. *SAGA: Journal of Technology and Information System*, 1(4):121–129, 2023.

[23] Alan R Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS quarterly*, pages 75–105, 2004.

[24] Md Shamim Hossain, Md Mahafuzur Rahman, Abu Eyaz Abresham, Asif Jaied Pranto, and Md Raisur Rahman. Ai and machine learning applications to enhance customer support. In *Handbook of Research on AI and Machine Learning Applications in Customer Support and Analytics*, pages 300–324. IGI Global, 2023.

[25] DataStax Inc. An llm agent reference architecture. 2024.

[26] Cheonsu Jeong. Generative ai service implementation using llm application architecture: based on rag model and langchain framework. *Journal of Intelligence and Information Systems*, 29(4):129–164, 2023.

[27] Ziwei Ji, YU Tiezheng, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating llm hallucination via self reflection. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[28] Jean-Marie John-Mathews. Some critical and ethical perspectives on the empirical turn of ai interpretability. *Technological Forecasting and Social Change*, 174:121209, 2022.

[29] Anis Koubaa. Gpt-4 vs. gpt-3.5: A concise showdown. 2023.

[30] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[31] Yuxi Li. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*, 2017.

[32] Lutkevich.B. Gpt-3 (generative pre-trained transformer 3). *TechTarget*, 2023. Accessed: 2024-02-21.

[33] D. Marr. Artificial intelligence—a personal view. *Artificial Intelligence*, 9(1):37–48, 1977.

[34] Fernando Martínez-Plumed, Lidia Contreras-Ochando, Cesar Ferri, José Hernández-Orallo, Meelis Kull, Nicolas Lachiche, María José Ramírez-Quintana, and Peter Flach. Crisp-dm twenty years later: From data mining processes to data science trajectories. *IEEE transactions on knowledge and data engineering*, 33(8):3048–3061, 2019.

[35] Mayerhofer, Noah and Nyström, Sandra. Designing a Machine Learning Application to Obtain Customer Insights in the Banking Domain, 2022. Student Paper.

[36] John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI Magazine*, 27(4):12, Dec. 2006.

[37] Keivalya Pandya and Mehfuza Holia. Automating customer service using langchain: Building custom open-source gpt chatbot for organizations. *arXiv preprint arXiv:2310.05421*, 2023.

[38] Nikhil Patel and Sandeep Trivedi. Leveraging predictive modeling, machine learning personalization, nlp customer support, and ai chatbots to increase customer loyalty. *Empirical Quests for Management Essences*, 3(3):1–24, 2020.

[39] Ehud Reiter. A Structured Review of the Validity of BLEU. *Computational Linguistics*, 44(3):393–401, 09 2018.

[40] Per Runeson, Emelie Engström, and Margaret-Anne Storey. *The Design Science Paradigm as a Frame for Empirical Software Engineering*, pages 127–147. Springer, Germany, 2020.

[41] Michael Schmidt. Clarifying the uses of artificial intelligence in the enterprise. *TechCrunch*, May 2016.

[42] Umair Shafique and Haseeb Qaiser. A comparative study of data mining process models (kdd, crisp-dm and semma). *International Journal of Innovation and Scientific Research*, 12(1):217–222, 2014.

[43] Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17, 2023.

[44] Róbert Szilágyi and Mihály Tóth. Use of llm for smes, opportunities and challenges. *Journal of Agricultural Informatics*, 14(2), 2023.

[45] Oguzhan Topsakal and Tahir Cetin Akinci. Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In *International Conference on Applied Engineering and Natural Sciences*, volume 1, pages 1050–1056, 2023.

[46] Ivanov.I Wardini.J. Ai statistics: How many companies use ai in 2023? *TechJury*, 2023. Accessed: 2024-02-21.

[47] Peter Weber, Roland Gabriel, Thomas Lux, and Katharina Menke. *Information Management*, pages 247–279. Springer Fachmedien Wiesbaden, Wiesbaden, 2022.

[48] Jiaheng Wei, Yuanshun Yao, Jean-Francois Ton, Hongyi Guo, Andrew Estornell, and Yang Liu. Measuring and reducing llm hallucination without gold-standard answers via expertise-weighting. *arXiv preprint arXiv:2402.10412*, 2024.

[49] Marco A Wiering and Martijn Van Otterlo. Reinforcement learning. *Adaptation, learning, and optimization*, 12(3):729, 2012.

[50] Rüdiger Wirth and Jochen Hipp. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39. Manchester, 2000.

[51] Sam Wiseman, Stuart M Shieber, and Alexander M Rush. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*, 2017.

[52] Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*, 2023.

[53] Adam Zaki. 67% of companies continue to adopt ai slowly, report. *CFO.com*, 2023. Accessed: 2024-02-21.

[54] Min Zhang and Juntao Li. A commentary of gpt-3 in mit technology review 2021. *Fundamental Research*, 1(6):831–833, 2021.

Appendices

Appendix A
Interview Guide

The following questions were used as a checklist of questions to be answered during the the focused interview (see Section 3.1.1). The purpose of the questions was to gain insights into the business operations and documentation practices within the case company (see Section 4.4).

- Can you describe the typical process for creating documentation in your department? What steps are involved in the process?
- What are the most common issues or difficulties you encounter when managing documentation? Are there specific tasks that are particularly demanding?
- How much time do you and your team typically spend on creating a document?
- What tools or software do you currently use for documentation? Do you find them efficient, or do they have limitations that hinder your work?
- How do you ensure the quality and consistency of documentation? Are there existing processes for review and validation, and do they work effectively?
- If you could change or improve any aspect of the documentation process, what would it be? What specific solutions or technologies would you find most helpful?
- Based on your experience, do you have any suggestions for improving the documentation handling process? What changes would make your work more efficient and less prone to errors?

Appendix B
Feedback Form for Application Evaluation

The following questions were used to capture insights and specific feedback during the evaluation process (see Section 3.5).

- What are the main differences between the result from the M1-model and the M2-model?
- Would you say that any of the models produce on average a document that is better than the average document currently existing in the knowledge database? Why/why not?
- Have you noticed hallucination from the models? Does one model tend to hallucinate more than the other?
- Do you think this solution or something similar could reduce the workload for employees? If so, which employees would benefit the most?
- Do you think it is possible to improve the quality of documents in the knowledge database by using the M1-model to write documents?
- Did you notice any specific type of query or document that the M1-model excels at?
- Did the M1-model perform better when given context?
- Do you think it is possible to improve the quality of documents in the knowledge database by using the M2-model to write documents?
- Did you notice any specific type of query or document that the M2-model excels at?
- Did the M2-model perform better when given context?

Appendix C
Prompts

The following prompts were analysed and assessed to determine how they yield factually accurate and linguistically sound results (see Section 4.5).

- **Prompt 1:**
  No text prompt was given {$context$}

- **Prompt 2:**
  You are a writer of documents for (the case company). Write a document with the following information: {$context$}

- **Prompt 3:**
  You are a writer of documents for (the case company). The documents you write should be based on the provided facts and formatted to comply with the company's tone of voice and content guidelines.

  {$context$}

  {$chatHistory$}

  {$instruction$}

- **Prompt 4:**
  You are an author of documents for (the case company's) Swedish internal documentation. The documents consist of four parts: Title, Description, Content, and Internal Content. The title is phrased as a question, the Description contains a brief summary of the document's key points, the Content section contains general information, and the Internal Content section contains information intended only for Customer Support to read. You will be provided with content for these four parts. Write a document based on this and ensure that the document aligns with (the case company's) tone of voice and guidelines.

  {$context$}

{*searchResult*}

{*chatHistory*}

{*instruction*}

- **Prompt 5:**
  You are an author of documents for (the case company's) Swedish internal documentation. The documents you write must be based on the facts provided and formatted to conform to (the case company's) tone of voice and content guidelines.

  Divide all documents into the following parts:

  - Title

  - Short description

  - Content

  - Internal content

  Answer the question based only on the following context and chat history:

  {*context*}

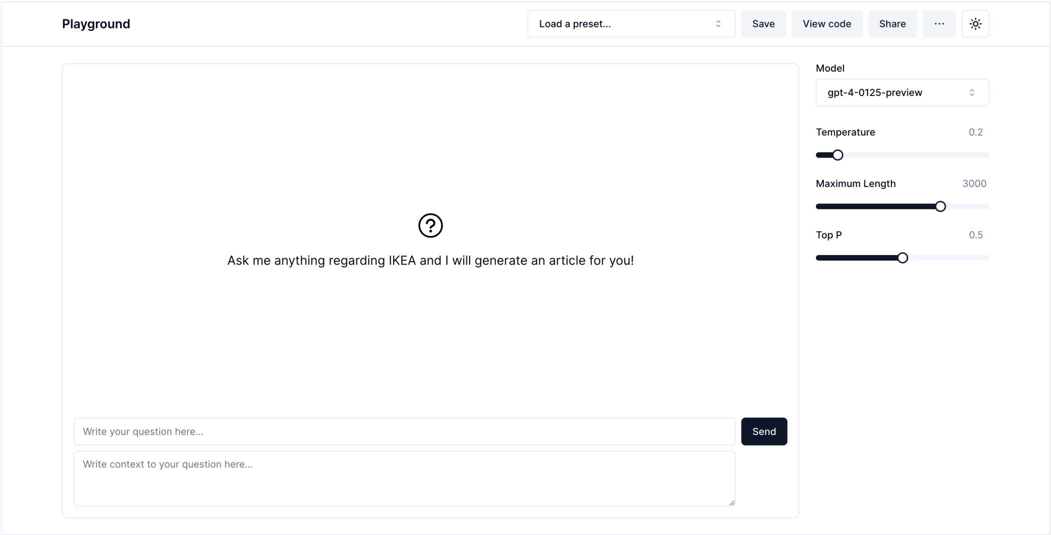
  The search result comes from the internet. Prioritise the context above the search results if the information is contradictory.

  {*searchResult*} {*chatHistory*}

  Instructions: {*instruction*}

Appendix D

Illustrations



Playground

Model: gpt-4-0125-preview

Temperature 0.2

Maximum Length 3000

Top P 0.5

Ask me anything regarding IKEA and I will generate an article for you!

Write your question here...

Write context to your question here...

Figure D.1: Visualisation of the application interface.

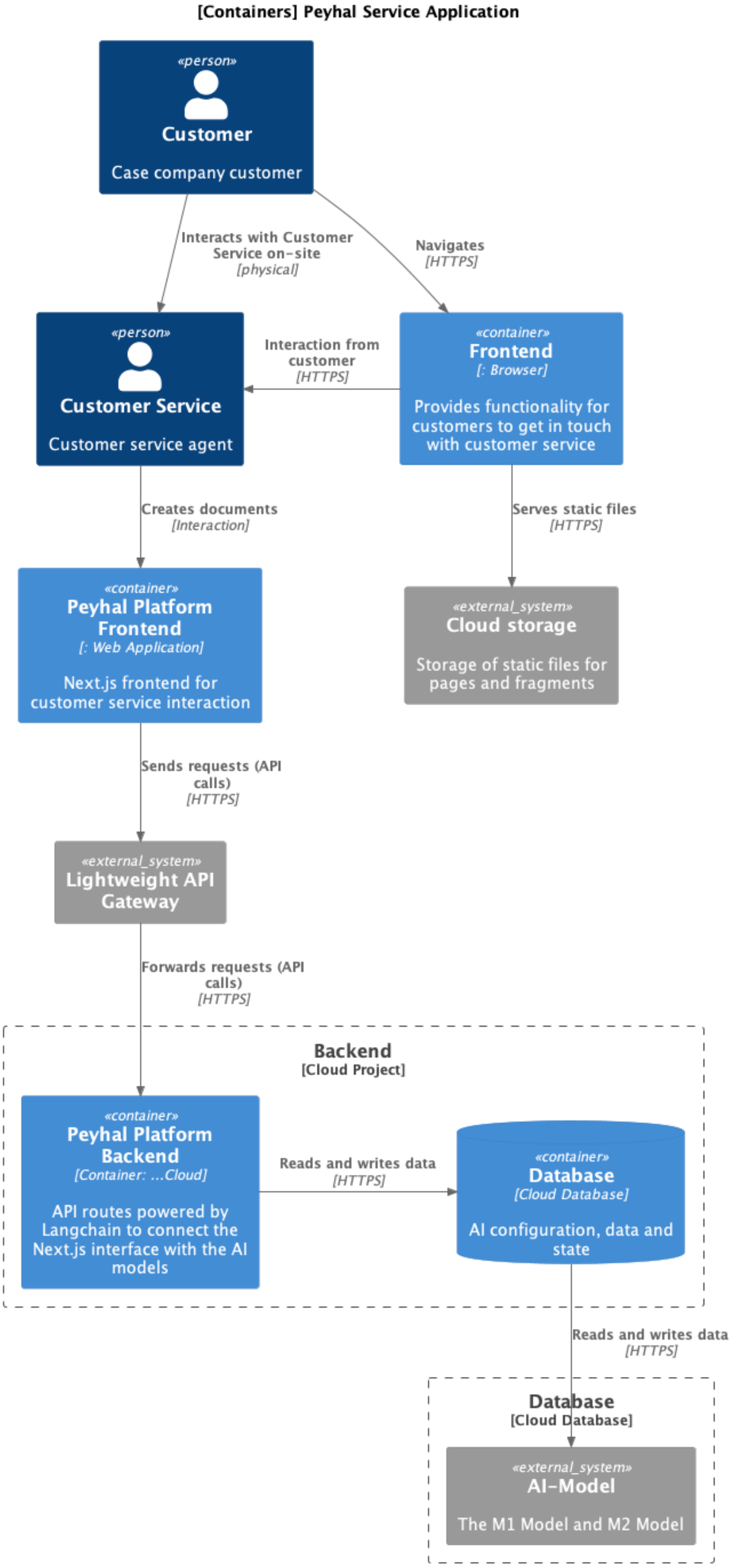


Figure D.2: Container Architecture and System Integration. This diagram shows how the PeyHal platform interact with the broader system, highlighting component relationships and data flows.

INSTITUTIONEN FÖR DATAVETENSKAP | LUNDS TEKNISKA HÖGSKOLA | PRESENTERAD 2024-05-30

EXAMENSARBETE Exploring the Potential of Generative AI for Corporate Documentation Management
STUDENTER Oscar Peyron, Oskar Hallberg
HANDLEDARE Elizabeth Bjarnason (LTH)
EXAMINATOR Alma Orucevic-Alagic (LTH)

AI som stöd för storskalig hantering av dokumentation

POPULÄRVETENSKAPLIG SAMMANFATTNING **Oscar Peyron, Oskar Hallberg**

I dagens informationssamhälle är företags förmåga att effektivt hantera dokumentation avgörande för att kunna stödja sina kunder och anställda. Detta arbete utforskar hur Artificiell Intelligens (AI) kan automatisera denna process och därmed skapa effektivare och mer pålitlig dokumentation.

För att företag effektivt ska kunna hjälpa sina kunder är det viktigt att de har välfungerande system för att dokumentera information om sina produkter, rutiner och processer. Mot denna bakgrund är det relevant att utforska nya tekniker som kan effektivisera hanteringen av dokumentation. Detta arbete utforskar hur användningen av AI kan automatisera skapandet av dokumentation inom ett multinationellt företag. Det företag vi har undersökt hanterar för närvarande sin dokumentation genom en komplex process som involverar flera olika intressenter, vilket resulterat i en rad olika utmaningar så som brist på skalbarhet, ineffektivitet och inkonsekvenser i dokumentens kvalitet.

För att adressera de identifierade problemen utvecklade vi en AI-baserad lösning som automatiserar skapandet av dokument. Vår lösning omfattar tekniker som anpassar och kompletterar AI-modeller för att generera dokumentationen som är korrekt och i enlighet med företagets språkmässiga riktlinjer. För att användare ska kunna interagera med de anpassade modellerna byggde vi även ett anpassat gränssnitt som fungerar likt en virtuell assistent.

Vår lösning har utvärderats genom att utvalda anställda inom företaget använt den för att skapa ny dokumentation och därefter bedömt innehållet utifrån ett antal parametrar. Lösning visade sig förbättra noggrannhet och minskade de inkonsekvenser som ofta förekommer i manuella processer. Den har även demonstrerat potential för att minska den tid och de resurser som krävs för att skapa dokumentation. Sammantaget uttryckte de anställda både intresse och optimism för framtida användning och vidareutveckling av lösningen. Detta understryker hur företag har möjlighet att förändra traditionella arbetsprocesser och frigöra resurser som kan användas i andra viktiga affärsområden genom användning av AI.

Vi tillämpade två olika metodiker (Design Science Research och CRISP-DM) när vi utforskade problemen med manuell hantering av dokumentation och när vi designade, implementerade och utvärderade vår lösning.

Vårt arbete lyfter fram både möjligheter och begränsningar med att använda AI inom dokumentationshantering. Genom att analysera de processer och resultat som presenteras, strävar vi efter att ge andra organisationer möjligheten att dra nytta av våra insikter och undersöka nya vägar för utveckling av AI.