Harnessing AI for Project Management Efficiency

Oscar Tyrberg and Tim Adenmark

DEPARTMENT OF DESIGN SCIENCES
FACULTY OF ENGINEERING LTH | LUND UNIVERSITY
2024

MASTER THESIS

Harnessing Artificial Intelligence for Project Management Efficiency

Developing and Implementing a Framework to Optimize Project Manager Tasks with AI Integration

Oscar Tyrberg and Tim Adenmark

Harnessing Artificial Intelligence for Project Management Efficiency

Developing and Implementing a Framework to Optimize Project Manager Tasks with AI Integration

Published by

Department of Design Sciences

Faculty of Engineering LTH, Lund University

P.O. Box 118, SE-221 00 Lund, Sweden

Subject: Innovation Engineering (INTM01)

Division: Division of Innovation Engineering, Department of Design Sciences, Faculty of Engineering LTH, Lund University

Supervisor: Lars Bengtsson

Examiner: Torben Schubert

Abstract

This thesis investigates and analyzes how to implement Artificial Intelligence (AI) into project management, addressing the process through a two-step approach. Initially, a framework is developed to identify and prioritize project management areas where AI can enhance operations. This framework was constructed based on a comprehensive literature review, adapting existing frameworks to the specific requirements of project management. It assesses tasks currently performed by project managers and ranks these tasks according to their potential for AI implementation. Subsequently, the thesis investigates the practical implementation of AI within these identified high-potential areas. Two AI solutions were developed as demonstrations; the first, a risk register utilizing a retrieval augmented generation (RAG) architecture, was evaluated to offer limited value. In contrast, the second demonstration, a budget tool designed to automate information extraction from PDF files, demonstrated significant potential. This tool successfully automated the extraction of information from various PDF structures provided by different suppliers, achieving a 98% accuracy rate for readable PDFs through techniques such as prompt engineering and fine-tuning. Furthermore, a business case analysis for the budget tool suggested a potential payback period of 0.36 years for deploying a fully functional application. The findings suggest that effective AI implementation in project management should begin with identification of tasks suitable for AI implementation. These tasks should then be prioritized based on financial, time, and risk implications, alongside the effort required for AI integration. The implementation process should foster collaboration between technical experts and domain specialists, embrace rapid iterative feedback, and initiate pilot demos for stakeholder evaluation prior to full-scale production. The thesis also concludes that successful AI deployment in organizations demands robust data management, data protection measures, comprehensive AI education for the workforce, and a culture that trusts but also critically evaluates AI solutions.

Keywords: Artificial intelligence, AI in project management, Generative AI, Prompt Engineering, Retrieval Augmented Generation, Fine Tuning.

Sammanfattning

Förevarande uppsats undersöker och analyserar hur artificiell intelligens (AI) kan implementeras i projektledning genom en process i två steg. Inledningsvis utvecklas ett ramverk för att identifiera och prioritera områden inom projektledning där AI kan förbättra det dagliga arbetet. Detta ramverk skapades baserat på en omfattande litteraturstudie, där befintliga ramverk anpassades till de specifika kraven inom projektledning. Ramverket bedömer de uppgifter som för närvarande utförs av projektledare, och rankar dem efter deras potential för att kunna förbättras genom AI. Vidare analyseras den praktiska implementeringen av AI inom dessa identifierade områden med hög potential. Två AI-lösningar utvecklades som demonstrationer; den första, ett riskverktyg som använder en arkitektur för *retrieval augmented generation* (RAG). Denna bedömdes erbjuda begränsat värde. I motsats visade den andra demonstrationen, ett budgetverktyg utformat för att automatisera informationsutvinning från olika PDF-format, betydande potential. Detta verktyg automatiserade extraktionen av information från olika PDF-strukturer som tillhandahålls av olika leverantörer. Den uppnådde en noggrannhet på 98% för läsbara PDF:er genom tekniker som *prompt engineering* och *fine tuning*. Vidare gjordes en investeringsbedömning för budgetverktyget som visade en potentiell återbetalningstid på 0,36 år för att bygga en fullt fungerande applikation. Resultaten tyder på att effektiv AI-implementering i projektledning bör börja med identifieringen av uppgifter med hög potential för AI. Dessa uppgifter bör sedan prioriteras baserat på ekonomiska, tidsmässiga och riskrelaterade implikationer, tillsammans med den tid och kompetens som krävs för att skapa AI-lösningen. Implementeringsprocessen kräver samarbete mellan tekniska experter och domänexperter, användning av snabb iterativ feedback och pilot-demonstrationer för utvärdering av lösningens potential. Utredningen resulterade i slutsatsen att framgångsrik AI-implementering i organisationer kräver robusta processer för datahantering, åtgärder för att skydda data, omfattande AI-utbildning för personalen och en kultur som litar på men också kritiskt utvärderar AI-lösningar.

Nyckelord: Artificiell intelligens, AI i projektledning, Generative AI, Prompt Engineering, Retrieval Augmented Generation, Fine Tuning,

Acknowledgments

This master's thesis was written during the spring of 2024 as the final part of the Master of Science program in Industrial Engineering and Management at the Faculty of Engineering at Lund University.

We would like to thank the project managers at the CAPEX implementation team at Tetra Pak for their engagement and support throughout this project. Special thanks to Andreas Wickman, our company supervisor, whose invaluable insights and support were instrumental to this thesis. We are equally grateful to Josefine Marklund Engström for her support. We also would like to thank our academic supervisor, Lars Bengtsson for the guidance and support throughout the process. Finally, we would also like to thank Jacob Tyrberg for his help in discussing various technical details throughout our work.

Lund, June 2024

Oscar Tyrberg, Tim Adenmark

Table of contents

List of acronyms and abbreviations

AI	artificial intelligence
CAPEX	capital expenditures
CRISP-DM	Cross-Industry Standard Process for Data Mining projects
GenAI	Generative Artificial Intelligence
GPT	Generative Pre-trained Transformer
LLM	large language model
ML	machine learning
NLG	natural language generation
NLP	natural language processing
PMBOK	project management body of knowledge
PMI	Project Management Institute
PoC	proof of concept
RAG	retrieval augmented generation
ROI	return on investment
WSJF	weighted shortest job first

1. Introduction

The introductory chapter provides a concise background that sets the context for the thesis and introduces the case company involved in the study. After the premise this section also articulates the purpose of the thesis, delineates the research questions that guide the investigation, and outlines the scope as well as the delimitations of the research.

1.1 Background

Artificial Intelligence (AI) has been a research topic since the 1950s, and machine learning has long been used to support decision-making and automating processes (Wang, 2019). However, it is first in recent years that AI has emerged as a revolutionizing technology with the use of Generative Artificial Intelligence (GenAI) and the release of large language models (LLMs) (Vaswani et al., 2017). With the advent of LLMs, companies that previously did not have the technological capabilities to utilize AI have started exploring its potential to improve their business processes, with as much as 79 percent of workers having had exposure to GenAI (Chen et al., 2023).

Project management has previously been one of the domains with the least use of AI due to the temporal nature of projects and the high reliability of human collaboration in projects (Hofmann et al., 2020). However, recent developments in AI technology have sparked a great interest in how project managers can augment and automate parts of their current work to allow for more time spent on value-adding activities (Nilsson et al., 2024). Given the strong interest among project managers and the high expectations for AI's impact on companies, exploring the best methods for integrating AI into project management is a relevant research topic. LLMs already have good general knowledge of project management, indicating a high future potential (Vakilzadeh et al., 2023). Recent technological developments and adaptations of LLMs have also opened the door to companies giving models access to their own data, something that greatly improves the usability of LLMs in organizations (Lewis et al., 2020; Shin et al., 2023).

There are several challenges related to implementing AI, both related to technical aspects but also related to the organization itself (Ångström et al., 2023). Organizations need to have well established policies on how to manage data quality, security, and availability (Mehri, 2023; Nagle et al., 2017). They also need to educate the personnel using the AI models to ensure the correct interpretation of results and avoid cultural challenges related to fear of AI taking jobs and a lack of trust in data (Gill & Kaur, 2023; Shrivastav, 2023; Arslan et al., 2021). These challenges are often not managed sufficiently, resulting in most AI implementation projects failing to provide the desired business value (Gartner, 2018).

This thesis aims to address both how and where project managers can avoid the pitfalls of implementing AI in their organization, as well as contribute to the interdisciplinary field of how recent AI technologies related to LLMs can be applied in organizations in general.

1.2 The Case Company

This thesis is conducted in close collaboration with Tetra Pak. Tetra Pak is an international manufacturing company that develops and produces packaging and processing solutions. Specifically, the thesis is done in collaboration with the capital expenditures (CAPEX) implementation team at the Industrial Base Engineering division at Tetra Pak. Tetra Pak Development & Technology Industrial Base Engineering is developing and implementing standardized manufacturing equipment and processes for converting factories. As a part of this process, the CAPEX implementation team supports the converting factories all over the world with installation projects.

The implementation of AI within Tetra Pak, and particularly within the CAPEX implementation team in a project management setting, presents challenges that need to be addressed to ensure the successful adoption and utilization of AI. Tetra Pak aims to explore areas within the project management process where AI can deliver substantial value while also identifying the necessary preparations and adaptations required for the deployment of AI technologies. Additionally, Tetra Pak wants to align efforts and facilitate the possible integration of AI into existing systems, ensuring that all personnel are informed of AI advancements.

1.3 Research Purpose and Questions

The purpose of this thesis is to investigate and analyze the implementation of AI in project management. The following research questions will be answered:

- How can teams identify and prioritize areas of the project management process that have the potential to be augmented by AI?
- How can AI be implemented in the identified areas to maximize project impact?
- What prerequisites are critical to ensure the success of AI implementation?

1.4 Delimitations and Scope

This thesis focuses on the exploration and application of AI within the context of the CAPEX implementation projects at Tetra Pak, focusing on the Project Manager's point of view. This means that potential AI implementations that include the work of other members of the project team, are not included in the report. The results include two demos of potential AI solutions in the process of the CAPEX implementation team at Tetra Pak. The purpose of these demos is to showcase the potential of the solution, and the steps of deploying the demos in production are not included.

1.5 Structure of Thesis

Chapter 1 - Introduction: A short background covering the premise of the thesis. An introduction to the case company. The purpose of the thesis with the following research questions are presented. Scope and delimitations are specified.

Chapter 2 - Theory: The academic foundation of the thesis is presented, including project management, AI, current research on how to implement AI in companies, and AI in project management with related challenges.

Chapter 3 - Method: The framework for how to identify, prioritize, and implement AI solutions in project management is presented and described. This is followed by the process of building the demo solutions.

Chapter 4 - Application of framework: The application of the framework described in Chapter 3 is applied to the processes of the CAPEX implementation team at Tetra Pak. The process of developing two demos is presented in detail.

Chapter 5 - Results and analysis: The results from the implementation of the derived AI solution are presented along with analysis of potential issues with the demos.

Chapter 6 - Discussion: Includes a summary of the findings along with a discussion concerning the limitations of the report along with potential future areas of research.

Chapter 7 - Conclusion: The final chapter concludes the report and answers the research questions given the results of the report.

2. Theory

The following chapter presents relevant and applicable literature for this thesis. It aims to create an understanding of key concepts, and to include relevant frameworks to help provide context for the subsequent methodology, result and discussion.

2.1 Intro to Project Management

A project is defined by the Project Management Institute (PMI) as a temporary endeavor, undertaken to create a unique product, service, or result (PMI, 2021). As a consequence of this definition, a project differentiates from ongoing, routine business operations based on two key characteristics: firstly, the uniqueness of a project and its outcome; secondly, the temporary nature of projects which always possess a defined start and conclusion (Oguz, 2022).

Building upon this description of a project, Bansal (2023) describes project management as "the process of leading the project team to achieve project objectives or complete project deliverables within the agreed time duration, allocated budget, and quality". In contrast, PMI (2021) defines it as "the application of knowledge, skills, tools, and techniques to project activities to meet the project requirements". Both definitions recognize that project management aims to achieve specific objectives or requirements of a project. This alignment in definition reflects a standardized way of viewing project management, acknowledging it as a disciplined approach to achieve project goals within a set of constraints and requirements.

2.1.1 Project Management Standard

PMI has been a leading advocate for the standardization of project management practices (ANSI, 2024). The Project Management Body of Knowledge (PMBOK), first published in 1987, provides a comprehensive framework released by the PMI that includes a range of processes considered best practices in the field and is one of the most referenced frameworks in project management (Takagi & Varajão, 2020).

The PMBOK Guide, as noted in its 7th edition by the PMI (2021), organizes its standard around knowledge areas, which are referred to as specialized fields commonly utilized in project management. These fields are described in Table 2.1 and are applied in the majority of projects. Specific project requirements may need additional knowledge areas. PMBOK offers guidelines and best practices for effective management. These disciplines are not standalone entities but are interlinked, each relying on the others to create a project management strategy.

Table 2.1: The ten different knowledge areas of PMI standard for project management (PMI, 2021)

Knowledge Area	*Short description*
Integration Management	Processes and activities to identify, define, combine, unify, and coordinate the various processes and project management activities
Scope Management	Processes and practices dedicated to precisely defining and regulating the scope of a project.
Schedule Management	Processes and practices to plan, develop, maintain, and control the project timeline to ensure the timely completion of the project.
Cost Management	Processes and activities involved in planning, estimating, budgeting, financing, funding, managing, and controlling costs of the project.
Quality Management	Processes aimed at integrating an organization's quality policies into the planning, management, and control of project and product quality.
Resource Management	Processes to identify, acquire, and manage all necessary resources to ensure the successful completion of a project.
Communications Management	Processes for timely and appropriate planning, collection, creation, distribution, management and control of project-related information.
Risk Management	Processes of conducting risk management planning, identification, analysis, response planning, and monitoring risk on a project.
Procurement Management	Processes essential for obtaining products, services, or outcomes from external sources.
Stakeholder Management	Processes to identify and manage entities with influence over project outcome.

2.2 Definitions and Overview of Artificial Intelligence

The definition of Artificial Intelligence (AI) has evolved over time, reflecting the rapid progression of the field. While the introduction of LLMs has brought AI to the attention of the general population, the term was first coined by John McCarthy in 1955. He described AI as "the science and engineering of making intelligent machines" (McCarthy et al., 1955). Using this definition, AI is machines that are capable of performing tasks that would typically require human intelligence. There has since been made a number of attempts at defining AI through different perspectives. One of the most recognized attempts was made by Russel and Norvig (2010) who further categorized four potential goals of creating AI systems. These are:

1. Systems that think like humans
2. Systems that act like humans
3. Systems that think rationally
4. Systems that act rationally

They thus made the distinction between systems that can merely imitate humans in the way they think and act and systems that can potentially surpass human intelligence. Another approach to classifying different types of AI systems in comparison to human intelligence widely used in the field is that of narrow, general, and superintelligent AI. These can be described as follows:

- **Narrow AI (ANI)**: A type of AI designed to perform a narrow task such as facial recognition or playing a game. A number of notable ANI systems have been produced, such as IBM's Deep Blue (Campbell, Hoane, & Hsu, 2002).
- **General AI (AGI)**: Generally described as AI systems that show human-like performance in general environments (Morris et al., 2024). The consensus among experts is that there is currently no active AGI today, while some say that modern applications (such as Chat-GPT) show AGI-like performance in some tasks (Bubeck et al., 2023).
- **Superintelligent AI (ASI)**: This represents an AI that surpasses human intelligence across a wide range of fields. The concept of ASI extends beyond current capabilities of AI and remains subject of much speculation and future research.

There are several terms that are frequently used in the context of AI, often interchangeably. See Figure 2.1 for an overview of the different types of AI. Machine learning (ML) is often used as a synonym for AI, but is, in fact, a subset of AI. Machine learning is the process of a machine to automatically learn and improve performance by analyzing large amounts of data. This is in contrast to

so-called expert systems that elicit human-like behavior through explicitly coded programs and large databases (Myers, 1986). There are a very large number of machine learning algorithms in use today, ranging from simple regression models to large neural networks with billions of parameters. Henceforth, machine learning will be referred to as the process of learning from data and making predictions based on that learning.

Neural networks are frequently discussed in AI and machine learning literature. These structures are inspired by the human brain and are commonly used in machine learning algorithms (IBM, n.d). They are applicable to problems of high complexity that require the analysis of very large amounts of data. Neural networks are composed of "layers", usually including an input layer, one or several hidden layers and an output layer. Deep learning is another commonly used term that is a subset of machine learning and is the naming of methods composed of neural networks with more than three hidden layers. Deep learning removes some of the need for human interaction in the modeling of data and is the basis of the advanced AI models used today (IBM, 2022).

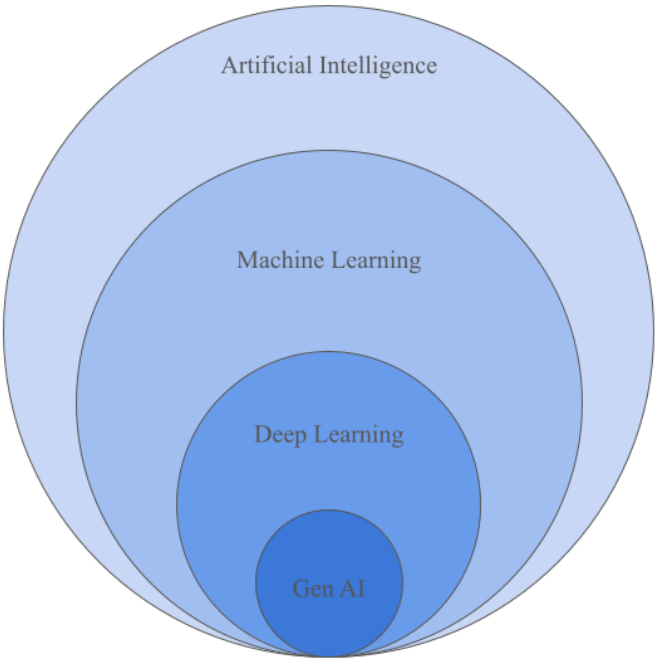

Figure 2.1: The different types of AI (Adapted from IBM, 2022)

2.2.1 Generative AI

Generative Artificial Intelligence (GenAI) encapsulates AI systems that can create content and data that mimic human-like behavior. GenAI has the ability to learn patterns, styles, or rules from existing data and use this information to generate new, unseen outputs. This is in contrast to other types of AI that are able to analyze data and make predictions, but are unable to generate new content. GenAI systems are built on large neural networks, usually with millions of parameters used to train the models. GenAI has proven applications in several domains, such as art creation, music composition and text generation (Goodfellow et al., 2014).

2.2.2 Large Language Models (LLMs)

Large Language Models (LLMs) are a subset of GenAI, primarily focused on processing, understanding, and generating human language. LLMs are trained on extensive corpuses of text data and are capable of performing a range of language-related tasks, including but not limited to, translation, summarization, question answering, and content creation. The architecture of LLMs is based on neural networks, particularly transformer models, which have revolutionized natural language processing (NLP). These models learn to predict and generate text by understanding the context and relationships between words in large datasets. The effectiveness of LLMs is directly related to the size of their training data and the complexity of their neural network architecture. They are known for their ability to generate coherent and contextually relevant text, making them useful in various applications such as chatbots, writing assistants, and automated content generation (Vaswani et al., 2017).

2.2.3 Generative Pre-trained Transformer (GPT)

The Generative Pre-trained Transformer (GPT) is a specific type of LLM developed by OpenAI. GPT models, particularly the latest versions like GPT-3 and GPT-4, stand out for their size, complexity, and wide range of capabilities. GPT-3, for instance, comprises 175 billion parameters (Brown et al., 2020).

2.2.4 Foundation Models

Training an LLM from scratch requires billions of data points and can take months to accomplish. The training of GPT-3 with its 175 billion parameters required ~400 billion tokens to complete and required several thousand petaflops/day in computing power during training (Brown et al., 2020). This is the equivalent of millions of everyday computers in computing power (Wikipedia, 2024). These

demands have several implications on the use of LLMs in company applications since the development and training of a custom LLM requires huge amounts of resources and time. The introduction of foundation models revolutionized this process by allowing the use of already trained models with billions of parameters for specific tasks (Bommasani et al., 2021). There are several types of foundation models, but the main type in use today is LLMs. Foundation models such as GPTs are task agnostic, meaning that they can be applied to a wide range of tasks by adjusting the model using a much smaller set of data than the original training data. Task specific foundation models have a wide range of applications, examples of which are in healthcare, law, and education (Bommasani et al., 2021). For LLMs, it is also possible to adapt the model to a specific task using prompts and very small data sets. In this instance, one does not actually adapt the model parameters to the specific task, but simply uses the ready made model as it is by giving it intelligent prompts (Strobelt, 2023).

2.2.4.1 Examples of foundation models today

There are a large number of foundation models available today, with a number of different structures and use cases. Most foundation models are LLMs, but there are models with other use cases such as image recognition and robotics (Bommasani et al., 2021). For the remainder of this paper, foundation models will refer to LLMs if nothing else is stated. Table 2.2 provides an overview of six of the most commonly used foundation models today and their model structure.

Table 2.2: Commonly used foundation models and their use case (OpenAI, 2023a; Google Research Team, 2024; OpenAI, 2024; Microsoft, 2021; Google, n.d; Huggingface, n.d)

Model name	*Company*	*Number of parameters*	*Short description*
GPT-4	OpenAI	Undisclosed (>175B)	Used for a variety of NLP tasks, including text generation, translation, summarization, question-answering, and more.
BERT	Google	340M	Primarily used for NLP tasks such as sentiment analysis, named entity recognition, and question-answering.
DALL-E 2	OpenAI	Undisclosed (>12B)	Used for generating images from textual descriptions.
ResNet	Microsoft	60M	Primarily used for image recognition and classification tasks.
PaLM	Google	540B	Designed for a wide range of NLP tasks, including language understanding, generation, translation, and question-answering.

T5	Google	11B	Designed to convert NLP problems into a text-to-text format. Used for translation, summarization, and question answering..

The evolution of foundation models is rapid and constantly ongoing, and new models are released frequently. Google recently announced their last multimodal model Gemini, that showed stronger performance than GPT-4 in several domains (Google Research Team, 2023). Meta recently released the open-source foundation model Llama3 which is freely available to use and performs in line with GPT-4 on most tasks (MetaAI, 2024). Further, OpenAI recently launched their latest foundation model GPT-4o with abilities in text, audio and video (OpenAI, 2024c).

As mentioned above, there are several ways to adjust foundation models to increase performance in specific tasks. The three most commonly used are prompt engineering, fine-tuning, and retrieval augmented generation (RAG), which are often used together in practical applications (Allard & Jarvis, 2024).

2.2.5 Prompt Engineering

To enhance the performance of foundation models, the first step should be using prompts to guide the model's behavior (Allard & Jarvis, 2024). Best practices for prompt engineering include (OpenAI, 2023b):

- **Writing clear instructions** such as asking the model to adopt a specific persona, using delimiters to specify parts of the instruction, specifying the steps needed to complete the task and providing examples of the desired output.
- **Providing reference text** could mean giving the model access to a specific text and asking it to answer by referencing the text.
- **Splitting complex tasks into subtasks** is effective given that most models have limits put on their input and output lengths. Long tasks like the summarization of a book can effectively be divided into shorter sub tasks such as summarizing each chapter on its own.
- **Giving the model time to think** could mean telling it to come to its own conclusion before providing an answer or telling it to go through a previous answer looking for errors.
- **Testing changes systematically** by making small changes and measuring their effect on the output in a way that allows fair comparison between methods.

Prompt engineering has several advantages compared to more advanced methods of improving foundation model performance. Prompting uses natural language and

does not require any writing of code which allows non-programmers to efficiently improve and evaluate model performance (Schmidt et al., 2023). It is also an effective way of establishing a baseline model performance in the early stages of development without requiring a lot of time or financial investment (Allard & Jarvis, 2024). However, prompt engineering is not enough if the model requires access to external data that was not used in the initial training. It is also often not effective enough when asking the model to format its output in a specific way (Allard & Jarvis, 2024).

2.2.6 Fine Tuning

Fine-tuning of LLMs means re-training a foundation model on a smaller, more specific data set than that which it was originally trained on. The data set used for fine-tuning should contain example input and desired output for the specific task. The model is then trained on the data, adjusting its parameters to fit the new training data (Microsoft, 2023). Fine-tuning is specifically well suited when a task requires high output precision or a specific output format or behavior (Allard & Jarvis, 2024). In specific tasks, fine-tuned, small LLMs have been shown to perform as well as, or better than, state-of-the-art models (Shin et al., 2023; Fatemi & Hu, 2023). Fine-tuning has also been used to drastically improve model performance in company-specific tasks where output precision was key (Allard & Jarvis, 2024). While OpenAI currently provides the most well-known platform for fine-tuning and access to models there are several other, both proprietary and open-source alternatives, to build fine-tuned LLMs. Examples include IBM's Watsonx platform and the open-source AI model library HuggingFace (IBM, 2024; HuggingFace, 2024).

2.2.7 Retrieval Augmented Generation

While fine-tuning has been shown to improve the performance of foundation models in specific settings, re-training a foundation model on large amounts of data is both time-consuming and costly (Lin et al., 2024). An alternative approach to fine-tuning is to use retrieval augmented generation (RAG) systems, first introduced by Lewis et al. (2020). The method encompasses using a pre-trained foundation model along with a Vector Database of external knowledge to provide additional context and facts. The method has been proven to outperform fine-tuned foundation models in knowledge-intensive tasks (Lewis et al., 2020).

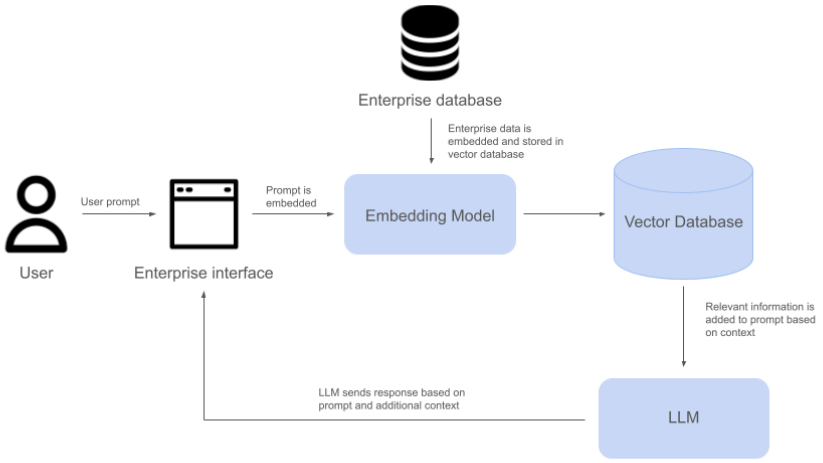

Figure 2.2: The RAG sequence architecture (adapted from Nvidia, 2023)

The process, as depicted in Figure 2.2, begins when a query is submitted to the LLM. This query is first processed by an Embedding Model, which translates the textual query into a numerical representation known as an embedding or vector. This transformation is needed to understand the query in computational terms. The generated embedding is then utilized to search for related embeddings within a Vector Database. This database is a repository of knowledge, potentially encompassing company-specific information or other relevant data that can augment the context of queries addressed to the model. Prior to storage in the Vector Database, this knowledge is similarly converted into numerical embeddings by the Embedding model; ensuring that all data within the database is in a standardized format. Upon identifying related embeddings in the Vector Database, the system enriches the initial query with this contextual information, creating an augmented query. This enhanced query, comprising the original prompt, the numerical representation of the query, and the additional context from the Vector Database, is then forwarded to the LLM. The LLM processes this comprehensive input to generate a response that is not only based on its pre-trained knowledge but also informed by the specific, relevant information retrieved from the Vector Database (Nvidia, 2023).

2.2.8 Evaluating LLM Performance

Evaluating the performance of LLMs presents a significant challenge, largely because various aspects make their assessment more complex compared to other machine learning models (Chang et al., 2023). LLMs often have billions of parameters (Brown et al., 2020), which makes it practically impossible to

determine exactly how a model has derived an answer. Where other machine learning models can be measured objectively by their error rates when making predictions or classifications, it is challenging to objectively measure the correctness of an answer produced by an LLM. There are currently over 50 benchmarks measuring general LLM performance. These benchmarks measure general NLP tasks such as sentiment analysis, text classification, factual correctness, and bias (Chang et al., 2023). Some of the most used evaluation methods are ChatBotArena and MT-bench, which incorporate human opinion in their evaluation metrics (Zheng et al., 2023). For downstream task performance, experts have evaluated LLM performance in several domains, including medicine, law, and engineering (Chang et al., 2023). For specific tasks in company settings, the choice of evaluation tools is often not straightforward and is often a combination of automatic evaluation and manual human evaluation. In theory, effective task-specific evaluation methods for companies should closely correlate with business outcomes, utilize a minimal number of distinct metrics, and be cost-efficient. In practice, this usually means combining automatic frameworks and human expert knowledge (Tobin, 2024). One example of an automatic benchmark suitable for company settings is the RAGAS framework, which is specifically designed to evaluate RAG applications. It evaluates model performance by evaluating the LLM and the information retrieval processes separately (Es et al., 2023). In summary, evaluating LLM performance in general is hard, and evaluating task-specific performance is harder. The literature on the subject is currently limited, but new methods are being developed rapidly.

2.3 Methodologies for AI Implementation in Organizations

The interest in utilizing AI in organizational settings has increased dramatically since the rise of GenAI and 55% of organizations state that they are currently using AI in some regard (Chui et al., 2023). However, a large number of AI projects fail. Gartner estimates that up to 85% of AI projects fail to produce accurate results (2018). The reasons for AI implementation failures vary, and there have been attempts at providing general purpose frameworks for AI implementation addressing the main reasons for implementation failures (Haefner et al., 2023). Several frameworks address specific parts, including the identification, prioritization, and implementation of AI applications. Some of the most commonly mentioned are the task-based approach for identifying projects presented by Autor et al. (2003), the weighted shortest job first (WSJF) approach for the prioritization of product development (Reinertsen, 2009), and the CRISP-DM framework for implementation of data mining projects (Wirth & Hipp, 2000).

2.3.1 Identification of AI Use Cases

The identification of possible AI use cases requires an understanding of the domains where AI can be implemented. To identify potential projects, Autor et al. (2003) propose a perspective on analyzing occupations, suggesting that each occupation can be viewed as a collection of tasks. This method underscores that certain tasks are inherently more responsive to technological implementation than others. To further elaborate on this task-based approach, Brynjolfsson et al. (2018), advocate for its application in determining the appropriateness of machine learning for specific tasks. They mean that based on a set of parameters particularly chosen for the domain, they can conclude how suitable the task is for implementing machine learning. The approach presented by Brynjolfsson et al. (2018) is specifically adapted for machine learning projects which differs from a more generalized AI approach. This means that these parameters or criteria need to be tweaked to suit other AI implementations, such as GenAI models. Brynjolfsson & Mitchell (2017) declare eight criteria that need to be fulfilled for tasks that are suitable for machine learning.

1. Function mapping between input and output
2. Large datasets
3. Clear feedback and goals
4. Simplicity in reasoning
5. Indifference to explanation of decisions
6. Tolerance for error
7. Stable phenomena
8. No requirement for physical dexterity

Many use cases of machine learning differ from the area or GenAI, and these criteria thus need to be adjusted when examining potential AI solutions. As described by Weisz et al. (2024), GenAI differs from machine learning in several ways. For example, GenAI does not require data labeling, does not include classification problems, and generally has lower explainability than traditional machine learning models. Other parameters that are not relevant in the case of machine learning are presented by Weisz et al. (2023). First, GenAI is by definition generative, which means its purpose is to produce artifacts as output rather than decisions, labels, classifications, and decision boundaries. Second, the outputs of a GenAI model are variable. Whereas machine learning aims for deterministic outcomes, GenAI systems may not produce the same output for a given input each time. In fact, by design, they can produce multiple and divergent outputs for a given input, some or all of which may be satisfactory to the user. Thus, it may be difficult for users to achieve replicable results when working with a GenAI application.

2.3.2 Prioritization of AI Use Cases

Generally, AI use cases in organizations should be prioritized based on their business impact and the complexity of implementation (Weber, Limmer, Weking, 2022; Brakemeier et al., 2024). A common approach to prioritize product development in agile environments that considers these two factors is the weighted shortest job first (WSJF) framework first introduced by Reinertsen (2009). The framework suggests prioritizing projects based on a score defined as $WSJF = \frac{Cost\ of\ delay}{Job\ duration}$. Reinertsen (2009) means that prioritizing the project with the highest WSJF score will lead to the best impact on overall results due to their relatively high costs of delay (impact) and their low job duration (complexity). In the WSJF framework, *cost of delay* is a proxy for a project business impact and is defined in Eq. (1) as

Cost of delay = User-business value + Time criticality + Risk reduction/Opportunity enablement. *(1)*

User-business value is described as the value the product will bring in terms of strategic and economic impact, as well as the perceived value of the end user. Time criticality measures how important it is to finish the product development in a given time frame, driven mainly by competing forces that could result in large negative impacts if the product is not finished on time. Risk reduction/opportunity enablement aims to measure the impact the product will have on the risks and opportunities of the business as a whole. For the second part of the WSJF framework, *job duration*, Reinertsen (2009) suggests using *job size* as a proxy since it is often hard to determine the exact time frame of new product development. *Job size* can be measured in several ways but should encapsulate the technical and organizational difficulties related to product development. Reinertsen (2009) suggests using a Fibonacci scale to determine the relative scores of each considered project for the parameters in the framework. The final *WSJF* score is calculated as shown in Eq. (2) below:

$$WSJF = \frac{Business\ value\ score + Time\ criticality\ score + Risk\ reduction\ score}{Job\ size\ score} \quad (2)$$

The product with the highest *WSJF* score is then prioritized for development first.

2.3.3 Implementation of AI Use Cases

There are several frameworks for how to best implement AI solutions, the industry standard being the CRISP-DM (Cross-Industry Standard Process for Data Mining projects) (Shröer et al., 2021; Dzhusupova et al., 2024; Bookkrantz et al., 2023). CRISP-DM, developed in the late 1990s by a consortium of technology

companies, provides a structured approach to planning and executing data mining projects (Wirth & Hipp, 2000). While still commonly used by industry professionals, it has received criticism for not addressing machine learning-specific tasks (Studer et al., 2021). Studer et al. introduced the CRISP-ML(Q) (Cross-Industry Standard Process for machine learning Applications with Quality Assurance) framework, extending on CRISP-DM and specifically addressing machine learning projects. The six steps of building machine learning models using the CRISP-ML(Q) framework are outlined in Table 2.3.

Table 2.3: The six steps of the CRISP-ML(Q) framework (Studer et al., 2021)

Phase	*Short description*
Business & Data Understanding	Define business objectives and translate to ML objectives. Collect and verify data quality. Assess project feasibility. Define success criteria.
Data Preparation	Produce data suitable to perform the modeling step. Includes selecting what data to use, cleaning the selected data and standardizing the data.
Modeling	A suitable ML model is selected and trained on the data from previous steps based on existing literature for similar problems.
Evaluation	Model performance should be measured on a validation data set. Results should be cross-validated by a domain expert.
Deployment	Model deployed either on cloud infrastructure or an embedded system depending on safety and performance needs.
Monitoring & Maintenance	The model is continuously monitored and maintained.

These steps of the CRISP-ML(Q) framework differ slightly from those of the standard CRISP-DM framework, mainly in the addition of the monitoring and maintenance step, and the collocation of the business and data understanding steps. Combining business and data understanding in the same phase highlights a key issue. It underscores the problem of keeping domain experts and data scientists separate during the initial planning of a machine learning project. It is considered best practice to merge these groups in the initial phase to ensure that business and technical objectives are aligned (Studer et al., 2021).

2.3.3.1 Defining success criteria

A key aspect of aligning technical and business goals is to determine what it means for a model to be successful, using criteria for success. The CRISP-ML(Q) framework highlights the importance of defining success criteria across three dimensions: business, machine learning, and economics (Studer et al., 2021). Success criteria from a business standpoint involve determining how the machine learning model will contribute to achieving specific business goals. This could include enhancing operational efficiency, increasing customer satisfaction, or

driving revenue growth. The economic success criteria focus on the financial implications of the machine learning project. This could involve considering the return on investment, cost savings, and overall impact on the company's bottom line. From a technical standpoint, success criteria relate to the performance of the machine learning model itself. For traditional machine learning models, this includes accuracy, precision, recall, robustness, and scalability (Studer et al., 2021). For GenAI applications, the technical success criteria could be set using an automatic evaluation benchmark or using human evaluators (Tobin, 2024). Incorporating these multifaceted success criteria ensures that machine learning projects are not only technically sound but also aligned with broader business and economic goals (Studer et al., 2021).

2.4 Applications of AI in Project Management

The application of AI in project management is a relatively new study. Depending on what is defined as AI and what is included in the project management processes, there are several different ways of approaching the subject. There are a number of challenges related to implementing AI in project management due to the temporal nature of a project and its unique results (PMI, 2021). Since learning from data is central to AI applications, the unique situations and lack of project-specific data contradict the use of AI in project management (Hofmann et al., 2020). Rather, a lot of decisions are based on the experience and knowledge of the project manager (PMI, 2021). However, in a report by PMI (Nilsson et al., 2024), investigating the use of AI in project management globally, 76% of respondents say they believe AI will have a profound impact on their profession in the next three years, indicating strong beliefs in AI among project managers. This is supported by Gartner (2019), who states that 80% of project management tasks will be automated by AI by 2030. While these numbers give an indication that project management will be highly impacted by AI in general, the amount of impact will likely vary between project types.

One common way of identifying the tasks related to project management processes is to refer to the PMBOK (PMI, 2021). One can then identify the implications of AI on those tasks (Fridgeirsson et al., 2021; Weng, 2023). According to PMI (2023), AI has the largest potential to automate tasks such as report generation and summarization of text. They also indicate that AI has the potential to assist human decisions in risk analysis, cost estimation, and the analysis of large datasets as depicted in Figure 2.3. This is in line with the findings of Fridgeirsson et al. (2021) which indicate that AI has the most impact where historical data is available, such as for forecasting, maintaining baselines, estimating costs, and monitoring risks. Fridgeirsson (2021) is also in line with PMI (2023) regarding the fact that AI is unlikely to be of use in processes requiring human collaboration, such as team

building and stakeholder management. According to Nilsson (2024b), numerous applications of AI in project management are already in production. Common use cases involve ChatGPT utilized for example in idea generation, document summarization, and task management. Other use cases described employ collaboration tools such as Trello and Jira to streamline project tracking and task prioritization. Additionally, Nilsson et al (2024b) mention instances of fully customizable solutions built from scratch, tailored to address specific challenges within the project management organization.

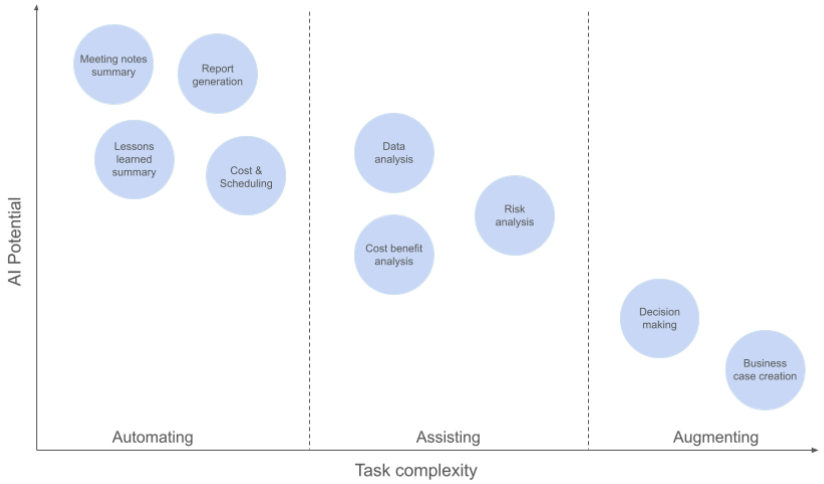

Figure 2.3: Potential of AI to support in project management processes (adapted from PMI, 2023)

Up until the launch of Chat-GPT in November 2022, the majority of literature was focused on the use of traditional machine learning methods in project management. Wang (2019) categorized the various forms of AI in project management as *Automation & Integration, Chatbot Assistant, Machine learning-based methods,* and *Autonomous project management.* His opinion was that the third category showed the most potential to aid project managers through various forecasting and estimation methods. This is supported by other studies as well, showing the benefits of machine learning in predicting project delays (Ahmad, 2015) and cost estimation (Tayefeh, 2020). The construction industry is currently the industry with the most applications of AI due to the high complexity of projects and relatively high amount of data available (Taboada et al., 2023). However, Wang's (2019) article was published three years before Chat-GPT's introduction. Since then, numerous reports have detailed the application of Chat-GPT and similar LLMs in project management, particularly for content creation and scheduling tasks.

Prieto et al. (2023) investigated the use of Chat-GPT in generating scheduling reports for construction projects and found great potential in the tool to automate time-consuming tasks. Weng (2023) outlines potential methods of implementing Chat-GPT across all tasks related to project management referred to in the PMBOK (PMI, 2021). The methods proposed are all related to report generation and automation and do not cover the use of traditional machine learning. Weng (2023) does not include any concrete examples of companies actually implementing said methods, so no conclusions can be drawn about the effectiveness of the approach.

As an indicator of LLMs potential use in project management, a study by Vakilzadeh et al. (2023) investigated the performance of GPT-4, GPT-3.5 and Google Bard on the Project Management Professional (PMP) test by PMI. They let each model complete the full test and reported that GPT-4 achieved an 87% score without any prompting aid. GPT-3.5 and Bard achieved 72% and 73% respectively. The results indicate that LLMs have a large potential to aid project managers with knowledge in project management without any specific training.

2.5 Challenges of AI Adoption

There are many challenges related to implementing AI solutions. As categorized by Ångström et al. (2023), the challenges are of technical, organizational and cultural nature. These challenges are not directly associated with project management, but as they are general on an organizational level, they become relevant in this setting as well. Finally, the environmental issues of building and leveraging large-scale AI systems are briefly mentioned. Figure 2.4 describes a general overview of the challenges mentioned in this section.

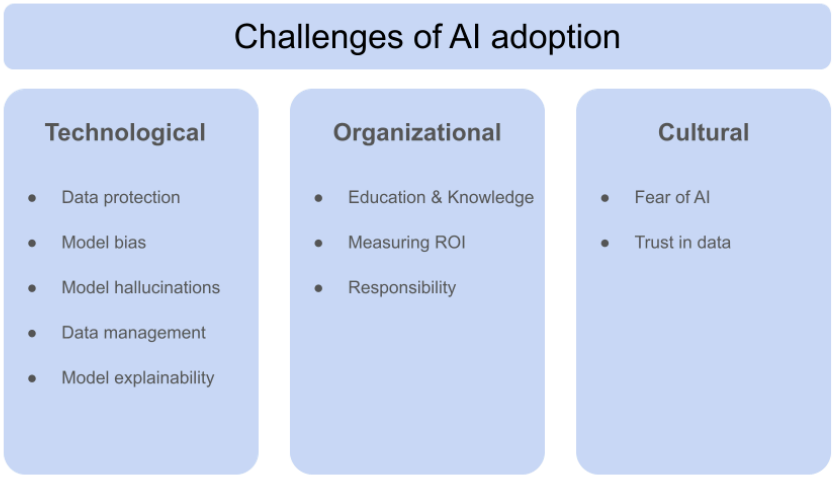

Figure 2.4: Challenges related to AI adoption

2.5.1 Technological Challenges

When the challenges stem from the AI technology or as a consequence of the technology, they are considered to be a technical challenge (Ångström et al., 2023). The acknowledged challenges in this section are data protection, bias, hallucinations, data management, and explainability.

2.5.1.1 Data protection

The challenge of data protection is a concern in the context of AI models. The safeguarding of sensitive and private information is a constant threat if the data is used in models or other software not run or owned by the organization. Both Fui-Hoon Nah et al. (2023) and Gill & Kaur (2023) identify issues and risks when it comes to data protection in AI models and GenAI models in particular, and describe the possibility of disclosing sensitive or private information if not used with caution.

An example of such a risk was observed in incidents involving ChatGPT and Bard AI where users could access other users' data (Sriram, 2023; Patel, 2023). This phenomenon is highlighted by Hitaj (2017) as a common issue for AI models that handle shared information, thus posing significant risks of data mishandling. In GenAI models, this issue is even more emphasized. As raised by CEDPO (2023), GenAI models have the potential to extract sensitive information and then republish the information in response to other unintended users. This means that data used to prompt an AI model is incorporated into the model, free for other users to retrieve from similar prompts. Smith et al. (2023) delve into mitigation

strategies where sanitizing your data from sensitive and private information before using it to train the model is a privacy-preserving approach. Another approach by Hitaj (2017) recommends training AI models locally with sensitive information, advocating for a practice where sensitive data does not leave the premises of the organization. This local training method is seen as the most secure way of preventing the external exposure of sensitive information, but may lead to limitations in model performance.

2.5.1.2 Bias

Ferrara (2023a) defines bias as systematic errors that occur in decision-making processes, leading to unfair outcomes. For AI models, this can arise from the data, the algorithms used, and how the human interprets and uses these results. Schwartz et al. (2022) say that these biases can be introduced by intent or inadvertently into the system, or they can emerge as the AI model is being used. There are many sources of biases in the context of GenAI, the most important ones being data-driven bias, automation bias, and feedback loop bias (Ferrera, 2023a).

Data-driven bias is biases embedded in the training data that manifest in all AI models (Dwivedi et al., 2023). Dwivedi et al. (2023) highlight that biases such as gender, race, or other stereotypes in training data are inevitably absorbed by the model trained on this data. Zhuo et al. (2023) discuss representation bias, illustrating how it can lead to the underrepresentation of certain groups. Consequently, the content generated by these AI models may be less relevant or biased against these groups. To mitigate data-driven bias, Ferrera (2023b) emphasizes human-in-the-loop approaches in GenAI. These strategies involve human oversight at various stages of development and deployment. For instance, humans play a crucial role in curating and annotating training data, ensuring balance, and reducing controversial content's influence.

Automation bias, as first defined by Mosier & Skitka (1998), is the tendency to use automation as a heuristic replacement for information seeking and processing. This bias manifests, according to Potaznik (2023), when humans disregard or fail to seek contradictory information, accepting computer-generated solutions as correct. This is not due to laziness or ignorance but because of high trust in computer output. Such bias is evident in various settings where reliance on automated systems can lead to significant errors. Park et al. (2019) address automation biases by proposing a strategy that encourages decision-makers to consider the solution before accepting AI predictions. This method involves allowing decision-makers to reflect more deeply and consider a broader range of information. before relying on the AI-generated solution.

Feedback loop bias occurs due to the cyclical process where AI outputs can reinforce or exacerbate existing biases. This is a cycle or a feedback loop where the output of the AI tool or system influences its future outputs, reinforcing biases over time (Ferrara, 2023c). Schwartz et al. (2022) discuss how such feedback

loops can lead to disparity amplification. This occurs when marginalized individuals or groups are less likely to engage with an AI system, resulting in training data that predominantly reflects the most frequent users. For instance, non-native English speakers might be less inclined to use a voice-enabled personal assistant. Consequently, the experiences of these groups are underrepresented, causing the AI system to deviate from its intended purpose or functionality. To address these issues, developers can implement specific policies or guardrails. As an example, modifications to the behavior of ChatGPT and Bing-AI have been designed to mitigate unintended toxic behaviors or prevent malicious use (Ferrara, 2023b). These measures are crucial to prevent the further reinforcement of biases and ensure that AI systems function effectively and ethically.

2.5.1.3 Hallucinations

In the application of AI models that contain Natural Language Generation (NLG), i.e. models that generate natural language output, a common risk is hallucinations in the output text (Ji et al., 2023). The most inclusive and standard definition of hallucinations in the context of AI models, as described by several sources (Ji et al., 2023; Fui-Hoon Nah, 2023), is the phenomenon of NLG models generating unfaithful or nonsensical content in the provided source content.

In the organizational context, Fui-Hoon Nah et al. (2023) critically examine this aspect of GenAI, pointing out that hallucinations pose a significant risk due to the possibility of spreading misinformation or fake information. They argue that it is not entirely safe to trust the information generated by these models without scrutiny and that extra caution is needed when employing GenAI in situations where the expertise and judgment of professionals are crucial. This is elaborated by Hiter (2023), who points out the risk of reliance on unreliable data for analytics and critical business decisions.

In response to these challenges, Gill & Kaur (2023) emphasize the critical need for continuous monitoring and adaptive management of AI systems to mitigate risk. Several other attempts to mitigate the occurrence of hallucinations have been made, including specifying which data the model should use to answer a question, such as by implementing RAG architecture (Lui et al., 2020).

2.5.1.4 Data management

Adopting successful AI models depends heavily on the management of data, a concept central to AI systems and their output quality (Mehri, 2023). A study conducted by the Harvard Business Review showed that only 3% of company data fulfill general data quality criteria, which leads to large issues when implementing AI (Nagle et al., 2017). There are several frameworks designed to measure the quality of data, some of the most commonly referenced include Data Quality Assessment (DQA), Total Data Quality Management (TDQM), and Comprehensive methodology for data quality management (CDQ) (Cichy & Rass, 2019). The frameworks vary in their practical implementations, but most agree

that measuring data quality is a combination of objective measures and subjective measures (Cichy & Rass, 2019). They also agree that measuring data quality differs depending on the type of data. The three types of data relevant to measuring quality in the context of AI applications are:

- Structured data: Includes all data stored in a table format with clearly defined rows and columns. Examples are data stored in Excel files and relational databases.
- Unstructured data: Includes data with no clear structure. Examples are text files, and PDF documents.
- Semistructured data: Includes data that is a mix of structured and unstructured data. Examples are emails and JSON files.

While structured data quality can usually be measured objectively, unstructured data often requires subjective measures (Cichy & Rass, 2019). Common objective data quality dimensions include error rates, accessibility, and security. These dimensions can be measured numerically and are often given a score between zero and one. Common subjective measures include believability, interpretability, and relevance. These dimensions are not scored numerically but are rated based on domain expert opinions (Pipino et al., 2003). Different frameworks introduce different dimensions of data quality. The most common are presented in Table 2.4.

Table 2.4 data quality: Five common ways of measuring data quality (Cichy & Rass, 2019)

Dimension	*Short description*
Completeness	The extent to which data is of sufficient breadth, depth and scope for the task at hand.
Accuracy	The extent to which data is correct, reliable and certified.
Timeliness	The extent to which the age of the data is appropriate for the task at hand.
Consistency	The extent to which data is presented in the same format and compatible with previous data.
Accessibility	The extent to which information is available, or easily and quickly retrievable

For structured data, these dimensions are objectively measured and can be defined and calculated numerically. For example, completeness can be measured as one minus the number of missing values in the data divided by the total amount of data (Pipino et al., 2003). While most frameworks suggest that unstructured data is best measured subjectively by domain experts, attempts have been made to quantify the quality of unstructured data (Kiefer, 2016; Taleb, 2019). These frameworks measure similar characteristics as those for structured data but focus on quantification using keyword extraction and programs designed specifically to

measure unstructured data quality using machine learning methods. Kiefer (2016) suggests measuring unstructured data based on its interpretability, relevancy, and accuracy. These dimensions are further split into eight different concrete measures, and the author provides examples of how they can be calculated numerically. While these methods show promise, Kiefer (2016) acknowledges that there is more research needed in the field. In summary, there is a need for collaboration between data experts and domain experts when measuring data quality, especially for unstructured data.

Another concern in the management of data is the collection, storage, and distribution of data across the organization. As companies become more advanced in their AI use, the need for a centralized approach to data management increases (Ångström et al., 2023). A lack of data management processes leads to the concept of data debt, meaning a large accumulation of unnecessary data in the organization that makes the implementation of AI slower and more expensive. Managing data debt is a combination of clear data quality practices and collaboration between the producers and users of data (Tran, 2023). This also includes collaboration with external parties, including suppliers and customers. Business critical data may often be supplied by external parties, which puts pressure on external collaboration and cross-organization workflows (Ångström et al., 2023). Another key aspect in data management highlighted by Ångström et al. (2023) and Tran (2023) is setting clear standards for who is responsible for the data. This includes assigning responsibility for when, how, and where data is stored, as well as who is responsible for measuring the quality. Finally, Ångström et al. (2023) highlight that collaboration between operational experts and data scientists is crucial to creating value using data. It is not enough to rely on data scientists to manage all data matters since operational experts possess the required knowledge to determine what data is critical for the business and what is not. It is thus important that even operational experts possess basic knowledge regarding data management.

2.5.1.5 Low explainability of AI solutions

Gill & Kaur (2023) identify explainability as a significant challenge for AI models, noting that results generated by these systems can be difficult to interpret and even more challenging to explain to decision-makers or stakeholders. The complexity of AI decision-making processes often leads to outputs that are not intuitively understandable, necessitating a need for clearer explanation mechanisms. Furthermore, the lack of explainability can significantly impact user trust and the adoption of AI systems. Users and stakeholders are more likely to trust AI decisions when the rationale behind these decisions is clear and understandable, especially in critical applications (Petkovic, 2023).

Lack of explainability is especially present in the context of LLMs that usually have billions of parameters and unclear internal mechanisms (Brown et al., 2020).

It is generally very hard to determine how a LLM generated an output, causing risks of ill decision-making when there is a lack of governance (Zhao et al., 2024). Zhao et al. (2024) provide an overview of attempts at improving LLM explainability. They outline several examples, including the use of "Chain of Thought" prompts that instruct the model to explain how it came to a conclusion. However, they conclude that it is hard to determine whether these explanations actually help humans determine the internal processes that lead to the decision. They also mention several issues with the current attempts at improving explainability in LLMs, such as a lack of ground truth in how the models work. These issues present a significant challenge to the implementation of LLMs in organizations and often lead to limitations in the adoption of such models (El Zini, Awad, 2023).

2.5.2 Organizational Challenges

There are a number of potential challenges with implementing AI from an organizational perspective. The main challenges are related to the general AI competence in the organization, issues when measuring return on investment (ROI), and responsibility claims (Shrivastav, 2023; Ångström et al., 2023; Haefner et al., 2023). Ideally, all of these should be addressed during the implementation process and require the input and collaboration of all involved parts of the organization (Ångström et al., 2023; Haefner et al., 2023).

2.5.2.1 Education and knowledge

General AI competency across the organization is one of the main hurdles when first starting work with AI (Shrivastav, 2023; Ångström et al., 2023). A study by Amazon Web Services (2023) showed that 75% of employers struggle to hire the necessary talent to implement AI, and 82% of employers struggle to implement AI training programs for current employees. Out of project managers, 65% state that they do not possess any basic knowledge about AI (Nilsson et al., 2024), indicating that there is a big gap in AI competency among project managers. Organizations lacking competency in the use and development of AI applications also have a harder time attracting and retaining AI talent, indicating that getting started early and upskilling employees is critical (Beauchene et al., 2023).

The issue of lack of AI knowledge is not confined to data scientists and tech-related roles but is an issue in all parts of organizations. Ångström et al. (2023) mention that misaligned expectations of AI results due to a lack of AI competency among managers is one of the main challenges mentioned by firms. Misaligned expectations go both ways, as low expectations can lead to a reluctance to test new solutions, and too high expectations can lead to issues of trust as solutions fail to meet managers' expected results (Shrivastav, 2023). Misaligned objectives for AI solutions is another issue that is caused by a lack of

understanding of AI in the organization. While a solution may be optimal from a high-level strategic perspective, it may not be optimal on a tactical level. These conflicting interests in the organization make the implementation of AI solutions harder and require collaboration across levels (Shrivastav, 2023).

In general, organizations need to invest heavily in upskilling the entire workforce in regard to AI (Beauchene et al., 2023). The main difference between organizations with successful AI operations and those without, is that successful firms have significantly advanced their employees skills and attitudes toward AI. These firms are less likely to experience a shortage of AI-related talent, less likely to experience issues due to fear of AI, and have a smaller generational gap regarding employee preparedness (Ångström et al., 2023). The best way to approach upskilling is through fast iterations of early proof of concepts and pilot projects to create awareness of AI potential. This can be supplemented with traditional education, training, and workshops to provide additional knowledge about the subject (Ångström et al., 2023).

2.5.2.2 Measuring return on AI investment

Implementing AI requires high upfront investment, especially for companies that lack the technical infrastructure and knowledge needed to support the development. This, in combination with that AI solutions generally do not produce direct financial value in early implementation, means that organizations with limited resources may be hesitant to implement AI (Enholm et al., 2022). As mentioned earlier in this article, measuring AI success can be done in several ways, including technical, financial, and business-related measures (Studer et al., 2021). While some AI solutions, such as recommendation engines or customer service chatbots, may provide direct financial benefits, others are long-term initiatives and may not be measurable using traditional KPIs. Instead, companies can use productivity measures to value the impact of AI solutions (Borges et al., 2021). Such measures could cover process efficiency, insight generation, and business process transformation. However, the literature does not clearly define how to assess the return on investment for AI solutions concerning these measures, generally recommending a case-by-case approach (Enholm et al., 2022). In summary, a lack of clear KPIs that measure AI solutions' impact on financial performance can be an issue when determining if a solution should be implemented. This is especially true for organizations that lack the resources to run multiple small pilot projects without capturing the value of each.

2.5.2.3 Responsibility

The adoption of AI in the organization raises a challenge regarding responsibility for the actions and decisions made by the AI systems. Responsibility and accountability are mentioned as critical factors to consider when implementing AI in organizations, both because the impact of malfunction can be detrimental and also because a lack of trust in AI can stall progress (Merhi, 2023). The lack of

policy regarding the responsibility and accountability of AI systems has already affected large companies. For example, Air Canada recently lost a case in court where an AI chatbot had wrongfully given customers a refund for their air tickets (Garcia, 2024). While the Air Canada case did not result in large sums, the implication is that companies considering using AI solutions need to think about who is held responsible for the decisions made by the AI. According to Solaiman (2023), in the case that an AI system harms or is connected to harming people, who or what is to be held accountable is still unclear.

Recently, the European Commission introduced a proposal for an AI act similar to the previously introduced GDPR act, setting a general policy for the use of AI in organizations. The act sets specific measures of AI risk, divided into solutions with minimal risk, high risk, and unacceptable risk. The act holds the organization fully responsible for the impact of the AI solution, and companies failing to comply may be penalized with up to 7% of worldwide annual turnover (European Commission, 2023). The AI act highlights the importance for organizations to create clear policies for how to manage AI usage and what solutions are built. This needs to be done across the whole organization and includes both high-level managers and employees (Blackman & Vasiliu-Feltes, 2024).

2.5.3 Cultural Challenges

Out of the challenges that organizations face when implementing AI, cultural challenges are some of the most prominent. The most common challenges related to company culture are employee's fear of AI replacing jobs as well as a reluctance toward data-driven decisions (Ångström et al., 2023).

2.5.3.1 Fear of AI

Implementing AI in organizations can create fear and resistance from personnel, for example, due to fears of new technologies disrupting well-established practices (Arslan et al., 2021). The views on whether this will be the case are conflicting. Dwivedi et al. (2023) suggest that AI could automate many low-skilled or repetitive jobs, potentially making certain skills obsolete. Zarifhonarvar (2023) quantifies this impact, estimating that 32.8% of the workforce might experience a complete replacement of their work by GenAI, while 36.5% may face a partial impact, leading to changes in some tasks. Among project managers, 76% expect AI to significantly impact the profession in the coming three years (Nilsson et al., 2024). Atkinson (2016) takes a slightly opposing view and says that while technology may disrupt the way we work, it is not likely to eliminate jobs completely. Independent of view, everything indicates significant changes across organizations and industries with AI adoption. While this shift might lead to more

creative or different roles in the workforce, there's a lingering concern about AI replacing human jobs.

To deal with the fear of AI, Ångström et al. (2023) propose taking a change management approach centralized around upskilling employees and creating fair expectations. Key initiatives mentioned are running workshops and specific training, promoting skills development in AI, and changing current work routines. To drive interest among employees, Ångström et al. (2023) suggest that proof of concepts (PoCs) and demos are central to proving the potential value of AI and increasing the pace of adoption in the organization.

2.5.3.2 Creating a data-driven culture

The next cultural challenge regards employees' attitudes toward making decisions based on data. Ångström et al. (2023) mention that a common issue when implementing AI is employees' skepticism towards the model results when they disprove their intuition. It is not uncommon that decision makers prefer an intuitive approach over recommendations from data, even when they would benefit from a more data-driven approach. This is commonly mentioned as an issue for organizations that aim to become more data-driven in general, where employees' attitudes towards data is the main driving factor for successfully becoming data-driven (Berntsson Svensson et al., 2020). That decisions are made based on "gut instinct" and feelings is a very common issue among companies that aspire to be data-driven (Berntsson Svensson et al., 2020). This is especially relevant for project management, where the nature of projects leads to a lack of data and where decisions commonly rely on the experience and instinct of the project manager (PMI, 2021). To solve the issue of intuitive decision-making, Berntsson Svensson et al. (2020) suggest that companies can replace subjective decisions with AI and machine learning methods where applicable. While this can work in some instances, there are several issues related to this, for example, as shown by the Air Canada incident (Garcia, 2024).

A common issue when building a data-driven culture within organizations is a lack of trust in data (Berntsson Svensson et al., 2020), which will also be reflected in the trust in AI solutions. Lack of trust in data can be reflected both in a lack of trust in the quality of input data as well as the trust in the output of an analytical model. Lack of trust in the input data may be a result of poor data management both internally and externally (Berntsson Svensson et al., 2020). As mentioned in previous Section 2.5.1.4, data management is a common issue in organizations and something that should reasonably lead to a healthy skepticism toward the quality of data (Tran, 2023; Ångström et al., 2023). However, it is also common to have a lack of trust in the output of analytical models, especially if the output contradicts the instinct of the decision-maker (Passi & Jackson, 2018). This could potentially lead to wrong decisions being made or analytical recommendations being ignored (Berntsson Svensson et al., 2020). Managing the balance of a healthy dose of skepticism and ignoring analytical results because they are counterintuitive is best

done by fostering an open culture around data, as well as educating decision-makers on the potential upsides and downsides of using analytical models (Passi & Jackson, 2018). This goes hand in hand with improving the explainability of models (Gill & Kaur, 2023), as well as educating the workforce in AI solutions (Beauchene et al., 2024).

2.5.4 Environmental Challenges of Implementing AI

The environmental impact of AI is a complex issue. Neslen (2021) highlights AI's potential to mitigate environmental disasters and enhance energy efficiency, but this comes with significant computational demands. This demand is explained by Strubell et al. (2019), who emphasize the high energy consumption of training large AI models. According to Statista, the energy consumption of training GPT-3 reached 1250 Megawatt hours, the equivalent of the monthly energy consumption of 1500 American households (Statista, n.d.; Nussey, 2023). GPT-4 is estimated to have used six times as much energy to train as GPT-3 (TRG Datacenters, 2023). This showcases that energy consumption is likely to become an important aspect when building large-scale foundation models in the future. To solve this, Patterson et al. (2021) highlight the importance of selecting green data centers for training AI and also call for transparency in reporting the environmental footprint of AI systems. The environmental impact of training LLMs from scratch is not necessarily directly relevant to organizations leveraging foundation models for enterprise use cases since the marginal energy consumption from API calls to foundation models will be minimal (Kaspersen, 2023). On a larger level, Bender et al. (2021) underscore the uneven impact on marginalized communities affected by AI. These groups are less likely to access AI's advantages due to economic and technical constraints, yet disproportionately bear the brunt of climate change's adverse effects. While the environmental effects of AI will likely not be a top priority for organizations, it is important to keep in mind the downstream effects of scaling the technology, as models will continue to require more energy as they become more advanced (Patterson, 2021).

3. Method

This thesis adopts a tailored framework to identify, prioritize, and implement AI solutions. The development of this framework is grounded in an approach that includes theoretical research, discussions with domain experts at Tetra Pak, and the adaptation of pre-existing frameworks. By integrating established methodologies with current academic literature, a framework is tailored to leverage AI technologies effectively in project management settings.

3.1 General Overview of Work

The work of this thesis was split into three phases. See Figure 3.1 for an outline of the phases. Phase one included a wide literature study focused on developing a framework that could be used to identify and prioritize what AI solutions should be implemented first in the processes of the CAPEX implementation team at Tetra Pak. Phase one also included research about the current use of AI in project management in general, as well as what challenges are related to the implementation of AI in project management. Phase one laid the foundation for the subsequent steps of the thesis, and the findings are presented in Section 3.2.

Phase two included applying the relevant frameworks from phase one to the processes of the CAPEX implementation team at Tetra Pak. This was done in close collaboration with Andreas Wickman, project manager within the team. The first part of this phase was the identification of areas with AI potential using the task-based approach proposed by Brynjolfsson & Mitchell (2017). This part was divided into several meetings where the authors and Andreas went through all tasks related to his work as a project manager and identified the tasks with the potential to be augmented by AI. The second part was done using the WSJF approach proposed by Reinertsen (2009). It was similarly done over several meetings, where all identified tasks were scored and ranked according to the framework.

Phase three included the development of two demos based on the prioritization from phase two. This phase was structured using the CRISP-ML(Q) framework proposed by Studer et al. (2021). It included creating business and data understanding related to the task, data preparation relevant to the demo as well as

building the demo. Each iteration of the demos was tested against domain experts at Tetra Pak and updated according to the feedback provided.

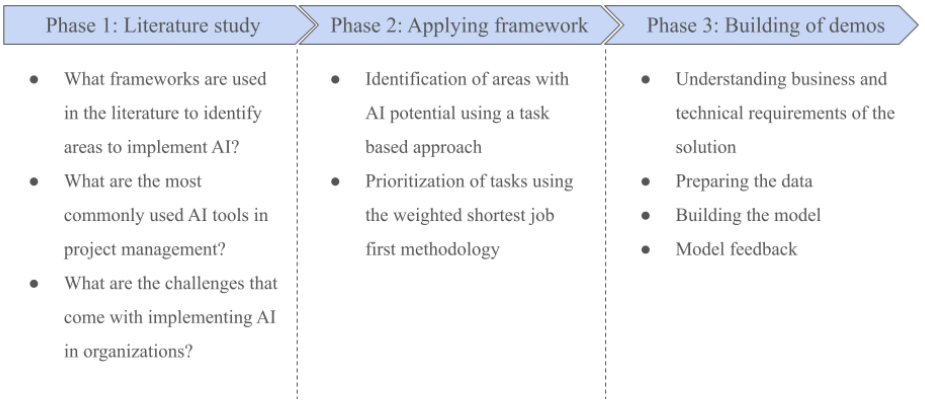

Figure 3.1: General process of work

3.2 Literature Study

In the literature study, scientific databases, IEEE Xplore, arXiv, Scopus, and ScienceDirect were used to retrieve relevant information. During the initial search, plenty of articles on AI were found, but specific information on project management, especially at the intersection between AI and project management, was lacking published research. To fill this gap, and to narrow the otherwise wide approach, other sources were needed. PMI is a nonprofit organization for project managers that has developed standards and best practices for the field. We chose to use their PMBOK, which is recognized by the American National Standards Institute (ANSI, 2024), to find standards for our work. PMI is a global organization, and we have been in contact with their Swedish division and also interviewed Marly Nilsson, project manager at PMI. She contributed with ideas and ensured that our methodology adhered to PMI's standards for project management.

During the period this thesis was written, several new studies on the subject were published. Notably, PMI released two reports: "Artificial Intelligence and Project Management: A Global Chapter-Led Survey" (Nilsson et al. 2024), and "Navigating AI in Project Management" (Nilsson et al. 2024b), the latter of which included and highlighted the framework developed and discussed in this thesis. These reports underscore the emerging nature of this topic and its current lack of a stable foundation in the academic literature.

The approach in this thesis was based on research questions aimed at identifying critical areas within the project management process. We then prioritized these areas based on their potential for AI application and developed methods to implement AI solutions to maximize positive impact. We divided the main questions into sub-questions, and for each, we identified best practices and frameworks through database searches and discussions with Tetra Pak. In many instances, specific research and frameworks directly relating to AI in project management were unavailable. Consequently, we adopted frameworks from similar contexts in other industries, applying these insights to the field of project management. This cross-industrial approach allowed conclusions and the possibility of tailored AI applications to project management.

AI tools were used sparingly throughout this thesis. In the initial phase, ChatGPT was used to get a basic understanding of the topics related to AI in project management. Gemini was also used for a number of specific research tasks, such as listing the most common foundation models and similar. Later, ChatGPT was used to structure the reference list by asking it to format the reference using the Harvard system of reference. However, this failed completely as ChatGPT included names of authors that were not included in the article. We thus had to redo this part manually later. For the most part, the work in this thesis has been completed in a traditional manner without the help of AI tools.

3.3 Identification and Prioritization of Tasks

The work of a project manager needs to be analyzed to identify possible areas where AI solutions can be implemented. As mentioned in Section 2.3.1, it is possible to segment the work of a project manager into different tasks, where all the tasks should cover everything done in the day-to-day work. This segmentation is accomplished in collaboration with a project manager, resulting in a comprehensive inventory of these tasks.

The subsequent step involves assessing the potential for an AI solution implementation within each task. The assessment is guided by a set of criteria derived from Section 2.3.1, drawing on insights from Brynjolfsson & Mitchell (2017), which includes an assessment of criteria for machine learning and Weisz et al. (2024) which highlights how GenAI differs from machine learning and how this may introduce new potential for AI solutions. Based on this, the following criteria were used to determine where it is suitable to implement AI in project management:

- The task has a lot of data related to it

- Knowledge about the task is thoroughly documented and can be accessed by an AI solution
- The task does not require a high degree of interpretability as to how conclusions have been made
- The effect of an error is not critical to project success

The subsequent stage involves ranking the identified tasks in terms of their implementation priority. Although each task has been recognized as a potential candidate for AI integration, determining the sequence for undertaking these AI projects is crucial. As outlined in Section 2.3.2, the weighted shortest job first (WSJF) method is an accepted strategy for prioritizing product development within agile frameworks (Reinertsen, 2009). Reinertsens approach was adapted for managing AI projects, aiming to prioritize tasks based on multiple factors. A customized version of this method is proposed, diverging from its traditional application in product development and instead tailored to the integration of AI solutions within project management. This adaptation is represented in Eq. (3), focusing on two primary variables: the magnitude of the business impact generated by the AI solution, the *project impact*, and the resource investment required for its implementation, the *task size*.

$$wsjf = \frac{Project\ impact}{Task\ size} \quad (3)$$

It is essential to define the two variables to effectively apply this prioritization framework. In a personal interview with domain expert Andreas Wickman at Tetra Pak, Eq. (4) and Eq. (5) defined the project's impact and the task's size.

$$Project\ impact = Financial\ impact + Time\ impact + Risk\ impact \quad (4)$$

$$Task\ size = Technical\ complexity + Data\ requirements \quad (5)$$

In an attempt to make this framework cover all potential areas, a final dimension of data availability is added to the equation, as shown in Eq. (6). Each dimension in the formula is explained in Table 3.1.

Table 3.1: Different dimensions that affect the prioritization score

Dimension	*Short description*
Financial impact	How important the task is for the financial end result including cost and revenue potential
Time impact	How much time is currently spent on the task by the project team
Risk impact	How much risk, both financial and technical, is associated with the task
Data availability	How much data that is currently available for the task

Technical complexity	Assessment of implementation complexity. Availability of off-the-shelf solutions leads to a low score, bespoke solutions lead to a high score.
Data requirements	How much and how good data is needed to implement the AI solution for the specific task

$$wsjf = \frac{Financial\ impact + \ Time\ impact + Risk\ impact + Data\ availability}{Technical\ complexity + Data\ requirements} \quad (6)$$

Reintsern (2009) discusses the challenges in quantifying dimensions but advocates for a relative approach by gathering domain experts to estimate these values. In this case, it is challenging to refer a quantitative measure to the dimensions (such as assigning an exact value to the financial impact a task has on the end result of a project). Instead, each dimension is graded in collaboration with a project manager, which gives merit to the final ranking of the framework. To make the model quantifiable, each dimension needs to be assigned a numerical value. In an attempt to make balance the interpretability of the application of the framework, each dimension is assigned a value of Low, Moderate, or High, which are in turn translated to values of 1, 2, and 3 respectively. These values are then used in Eq. 6 for the WSJF score. This approach allows the model to be quantified effectively.

3.4 Implementation of Prioritized Solutions

Following the identification and prioritization phases, the two tasks with the highest *WSJF* scores were used to develop demos for how AI can be used to augment the tasks. This step aligns with the principles outlined in Section 2.3.3 under the CRISP-ML(Q) framework, which provides a systematic approach to implementing AI solutions. Detailed discussions on the implementation for each of the AI solutions are described below with the CRISP-ML(Q) framework as a starting point.

4 Application of Frameworks

This section covers the applications of the task division approach to identifying tasks with AI potential, the weighted shortest job first framework for prioritizing AI solutions, and the CRISP-ML(Q) framework for implementing prioritized solutions.

4.1 Identification and Prioritization

This step includes the identification of tasks suitable for AI applications and the prioritization of identified tasks. Based on all tasks of a project manager at Tetra Pak, the tasks that completely lack AI potential are excluded from further evaluation based on the criteria outlined in Section 3.3. For example, the tasks of product verification, product validation, and agreement on high-level scope consist of interpersonal communication and relations and do, at this moment, lack AI potential at Tetra Pak. This is logical, given that they are not related to any form of documentation or data, they require a very high degree of explainability of conclusions, and the effect of an error in these steps would be critical to project success. In a personal interview with domain expert Andreas Wickman at Tetra Pak, they were thus excluded from the subsequent steps. See Table 4.1 for the full list of tasks.

Table 4.1: List of tasks for a project manager. Each task is described broadly rather than in detail.

Task	*Short description*
Verification & Validation	Oversee the completion of product verification and validation processes to ensure that the project meets standards.
Agreement on high-level scope	Securing consensus among key stakeholders on the project's overarching objectives and deliverables.
Identify parallel projects	Identify other ongoing projects within the company that are occurring simultaneously and may have intersections.

Assign project members	Allocating specific roles and responsibilities to form a project team
Open new project	Initiate a new project by setting it up in the company's project management system.
Review resource availability	Assess the current availability of personnel, equipment, and materials required.
Review and update budget	Review and update the budget with the latest financial information.
Predict travel cost	Estimate the expenses related to travel for project team members, including transportation and accommodation.
Predict installation and commissioning	Forecast the costs and time required for setting up and initializing the project's deliverables.
Predict contingency amount	Forecast the need for a contingency amount in the budget to cover potential risk-related costs
Create and update risk register	Compile a comprehensive list of potential project risks, their impact, and mitigation strategies to manage them effectively.
Review expert question list	Evaluate a compiled list of inquiries to be addressed by subject matter experts, ensuring all critical project questions are covered.
Create transport documentation	Prepare all necessary documents for the transportation of project materials or equipment, ensuring compliance with legal and safety standards.

Note: The information is designed to represent the common tasks but does not precisely mirror or replicate the practices at Tetra Pak.

Next, Eq. (6) was applied to all tasks with basic AI potential from Table 4.1. Each dimension in Table 3.1 was considered during a personal interview with project manager Andreas Wickman, who assigned a ranking of low, moderate, or high to each dimension. These rankings were then translated to numerical values of 1, 2, and 3 to be used in Eq. (6). Although this process is not an exact science, it relies on Andreas's expertise to provide a quantifiable method for generating a WSJF score for each task. This score effectively determines their priority order. The outcomes of this prioritization, including the WSJF scores and the ranked sequence of tasks, are compiled and displayed in Table 4.2. This systematic approach ensures a transparent and data-driven method for establishing the order in which AI solutions should be implemented, aligning project execution with strategic objectives.

As Table 4.2 below shows, the tasks of reviewing and updating the budget, as well as creating and updating risk registers, received the highest WSJF score. While these show the combined highest potential of implementation when weighing the effort required to implement each solution, it is also highly relevant to consider the tasks with a high *project impact* that received a low WSJF score due to a large *task size* and/ or low *data availability.* Examples of such tasks are to "create transport documentation", "predict contingency," and "predict travel costs". All of these were considered strong candidates for AI implementation and were all determined to have a high impact on project success either through large financial, time, or risk impact.

Creating transport documentation could have been automated using GenAI, where an LLM would be given access to relevant information regarding customs in different countries. However, it was determined that creating such a solution would be too complex for the scope of this thesis since it would have required collaboration with too many different stakeholders and a significant technical effort.

Predicting contingency and travel costs were both identified as key parts of managing project costs. Both of these could potentially have been achieved using traditional machine-learning approaches and historical project data. However, several issues were identified that made such an approach unfeasible. Currently, project data is stored in different databases and there is a lack of standards as to how data is stored, making the collection of data challenging. For contingency data, the data is not stored locally, and each factory has different ways of storing and managing contingency costs. Had the data been stored locally and easily accessible, these tasks would have received a significantly higher ranking.

Table 4.2: WSJF scores and priority ranking for the tasks of a project manager at Tetra Pak.

Task	*Financial impact*	*Time impact*	*Risk impact*	*Tech comp.*	*Data req.*	*Data availab.*	*WSJF Score*
Review and update budget	Mod	High	Low	Mod	Low	High	3,0
Create and update risk register	Mod	Mod	Mod	Mod	Low	High	3,0
Open new project	Low	Low	Low	Low	Low	Mod	2,5
Review resource availability	Low	Mod	Low	Low	Low	Low	2,5
Create transport	Mod	High	High	High	Mod	Low	1,8

documentation							
Predict contingency	High	Low	High	Mod	High	Low	1,6
Assign Project members	Mod	High	Mod	Mod	High	Low	1,6
Predict travel cost	Mod	Mod	Mod	High	Mod	Mod	1,6
Predict installation and comissioning	Mod	Mod	Mod	High	Mod	Mod	1,6
Review expert question list	Mod	High	Low	High	High	Mod	1,3

4.2 Process for Building Demo Solutions

This section covers the steps of applying the CRISP-ML(Q) framework to the two identified tasks with the highest priority scores.

4.2.1 Implementation of Demo - Review and Update Budget

Building upon the findings from previous analysis, it was determined that the process of reviewing and updating the budget is a task that stands to have potential benefit from the integration of AI solutions. To initiate the use of AI in the budget processes, an initial PoC solution was built using the CRISP-ML(Q) framework.

4.2.1.1 Business understanding

The first phase of the process builds upon the previous analysis that this has the potential of a successful AI solution. Currently, quotations from different suppliers have been received in PDF format with different layouts and styles. The different layouts make it hard to extract information in a standardized way, leading to manual extraction of information. Each project has quotations from several dozen suppliers which makes this a time-consuming task. The current workflow consists of receiving quotations and then manually extracting the needed information and manually inserting it into an Excel as described in Figure 4.1.

Figure 4.1: The current workflow.

Initial estimations showed that the project team currently spends approximately 4000 hours per year on this task. To automate this process would lead to significant productivity gains and make time for more value-adding activities. Initially, the proposed way of resolving the issue was to use an AI model to extract the relevant information from each quotation. The proposed workflow is described in Figure 4.2.

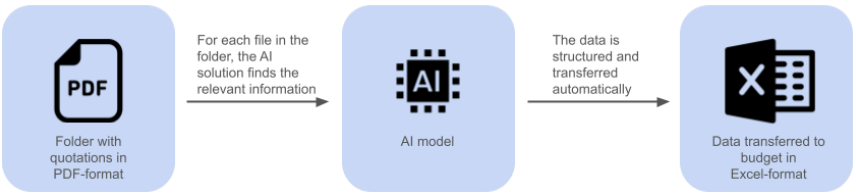

Figure 4.2: The possible workflow with an AI solution.

4.2.1.1.1 Calculating ROI for the solution

To give an understanding of the business impact of the solution, the potential net present value (NPV) and the potential payback period of the solution were calculated. To make the calculation, data on the costs of running the model were collected from OpenAI (2024), along with the man-hour costs of building the solution which are based on internal discussions with stakeholders at Tetra Pak. The potential cost savings of implementing the solution rely on full utilization of the model by the project managers. In extension, this assumption means that all the time saved by the model will result in created value, something that is not necessarily true. It is a somewhat simplified assumption, and cost savings would likely start lower and increase over time as utilization of the AI model would improve. Thus, the terms *potential cost savings, potential net present value* and *potential payback period* are used in the description of the financial model to signal the uncertainty of its results. A less optimistic scenario of 50% model utilization is also included for reference in the results. General assumptions made in the financial model are:

- Total potential cost savings stay equal for the entire period of the financial model
- Man-hour costs are kept equal from year 1 and onwards
- Development costs are kept equal from year 1 onwards
- Runtime costs are kept equal for all years
- Any infrastructure costs related to hosting the solution are absorbed by the larger organization and not included in the financial model
- A 10% discount rate is used for all calculations

4.2.1.2 Data understanding

The budget is a document in XLSX format (Excel) where cost information from quotations is stored. Today, the majority of information that is to be inserted in the budget comes from quotations that are documents in PDF format. Data needs to be transferred from unstructured PDFs to structured XLSX-files. Today, this is a manual task, but with the help of an AI solution, this may be solved automatically if the model can understand the unstructured PDF by extracting the correct data.

4.2.1.3 Data preparation

PDF is an efficient way to store a visual representation of a document, but the format is hard to work with if the goal is to extract structured parts of the text (Budhiraja, 2018). In this use case, the PDF documents cannot be processed directly by an AI solution, so the data needs to be extracted from each file. Different software can be used to extract the information. In an evaluation by Wiechork & Charão (2021), the usage of PDFminer was evaluated to be a working solution for extracting text. PDFminer extracts the text directly from the source code of the PDF (Marsman, 2024). In the demo, this was used to extract the information.

The next step in preparing the data was to clean the extracted text data. Irrelevant information, such as names and addresses, was changed to mock data or cleaned using a Python script.

The final step was to use the extracted and cleaned semi-colon-separated string and insert it into Excel. For this process, a Python script was used that consists of two main functions to handle and integrate data from a semi-colon-separated string into an XLSX file using the Pandas library. The first function transforms a string into a Pandas DataFrame and assigns predefined column names to the DataFrame. The second function inserts this DataFrame into Excel. It first attempts to read existing data from a specific sheet, combines it with the new data, and writes the result back.

4.2.1.4 Modeling

Based on Section 2.2.1, the use of GenAI should be a viable means to solve this use case. Since the quotations differ between suppliers and are highly unstructured, other, perhaps simpler, models would likely struggle to extract the relevant information. The need to form a high-level understanding of the quotation to draw conclusions on what information should be extracted is something that only GenAI can currently manage. A concrete example of GenAI used in this context is given by Ramachandran et al. (2024), who describes the possibility of automating data management processes using GenAI.

As noted in Section 2.2.4, a foundation model may need to be adjusted for the particular use case. In this scenario, the task requires high output precision, and the format needs to be able to be transferred to Excel. To achieve this behavior, a combination of prompt engineering and fine-tuning were deemed the most likely methods of adjusting the foundation models to the use case. The modeling was done in several iterations, with each iteration evaluated for its accuracy of results, as presented in Section 5. All iterations are presented in the following sections.

4.2.1.4.1 Attempt 1 - forming a baseline model

In the initial attempt, the foundation model GPT-3.5-Turbo was used due to its robust performance and free availability (OpenAI, 2023). A simple prompt was crafted to extract the relevant information and form a baseline performance of the model. This prompt was specifically utilized to capture the precise data fields required for transformation from the quotation. For the initial testing, only one supplier's quotations were used to build and test the model. The overall architecture of the first two attempts is presented in Figure 4.3. The prompt for attempt 1 is presented below:

- Prompt 1.0:
  The Quotation contains information about items that need to be extracted. The task is to extract information from the Quotation and output it as CSV. The CSV should have the format as follows where each row must contain 12 elements.

  [insert quotation number];[insert supplier];[insert item ID];[insert item description];[insert quotation date];[insert period of validity];[insert price];[insert discount];[insert unit price after discount];[insert currency];[insert quantity];;;

  """ Extracted Text"""

The outcome of this prompt was encouraging, though the output demonstrated variability across different model runs. This inconsistency may be attributed to technical challenges such as hallucinations and the inherent variability of GenAI models, as discussed in Section 2.5.1.3 and Section 2.2.1.

Figure 4.3: Overall architecture of the initial solution

4.2.1.4.2 Attempt 2 - Improved prompt

To enhance the accuracy, the prompt underwent multiple revisions to more closely align with the desired result. Key modifications included specifying the output content and format, altering the desired content, and incorporating specialized instructions tailored to the task at hand. The revised prompt is presented below:

- Prompt 2.0:
  The Quotation contains information about items that need to be extracted. The task is to extract information from the Quotation and output it as a CSV. If the information is not available, insert N/A. If there are several items in the Quotation, please provide the information for all items and add them to the CSV output as follows
  Each item should be separated by ;;; and a new line and the columns are separated by ;.
  The CSV output should have the format as follows where each row MUST contain 12 elements & should not be surrounded by "" & and you should not include the column names in the output:

  [insert quotation number];[insert supplier];[insert item ID];[insert item description];[insert date in format: YY/MM/DD];[insert the latest date in the validity period in format YY/MM/DD];[insert original price per item];[insert discount, if no discount found, set to 0%];[insert price per item after discount, if you don't find any discount, set this to the same as original price per item];[insert currency in format ISO 4217 (e.g. USD, EUR, SEK etc.)];[insert quantity];[insert category based on below];;;

  Category should be:

1. Spare parts if spare parts are mentioned in the item description.
2. Installation if installation is mentioned in the item description.
3. "Material 0010" if not installation or spare parts.

""" Extracted Text """

The results obtained from the revised prompt demonstrated a notable enhancement in accuracy, though still far from the desired output.

4.2.1.4.3 Attempt 3 -Introducing more advanced models

To further refine the model, the subsequent step involved fine-tuning the GPT-3.5-Turbo foundation model, as well as testing the previous prompt using the more advanced foundation model GPT-4. As detailed in Section 2.2.6, fine-tuning is particularly advantageous for models requiring high output precision or specific output formats, both of which are crucial for this task. To fine-tune GPT-3.5-Turbo, the OpenAI user interface was used. There, examples of ten quotations from the same supplier and the corresponding ideal output, along with ten synthetic quotations and their corresponding ideal outputs, were uploaded. The synthetic training samples were created using GPT-4 with one of the real quotations as a source of truth, see Appendix A for an example of a synthetic training sample. The model was then fine-tuned in the OpenAI interface on the additional data, with the assumption that it would be better at identifying the correct information from the quotations.

Both the fine-tuned GPT-3.5-Turbo and GPT-4 showed significant improvements compared to the previous models. For the single supplier, the performance of the fine-tuned model slightly beat that of GPT-4, showing that a smaller fine-tuned model can beat a larger, more advanced version, in line with the findings of Shin et al. (2023). These results were presented to key stakeholders at Tetra Pak, and the decision was made to further improve the model to make it usable for several different suppliers.

4.2.1.4.4 Attempt 4 - Applying the models to multiple suppliers

In Section 4.2.1.4.3, it was made clear that the initial PoC could handle quotations from a single supplier, leading to a preliminary acceptance of the approach. However, the challenge of generalizing the solution to accommodate multiple suppliers remained unresolved. To move forward, a closed and completed project with quotations from 20 different suppliers was used to establish an application.

The initial approach involved utilizing the fine-tuned model together with the highest-scoring prompt from previous experiments. However, this approach yielded suboptimal results when the model was presented with quotations from suppliers other than the one it was fine-tuned on. Next, the model based on GPT-4 was applied to the full project quotations, which gave a significantly better score

than that of the fine-tuned model, even though the accuracy was still lower than that of the models applied to a single supplier.

4.2.1.4.5 Attempt 5 - Building individual prompts for each supplier

An issue with applying one single model and prompt to several suppliers was that the different suppliers had widely different structures of quotations and used different language for the same things. This raised the idea to introduce individual prompts for each supplier and let the model identify which prompt to use based on which supplier the quotation belonged to. This could then also be extended to fine-tuning separate models for each supplier if needed. The overall structure of the refined model is presented in figure 4.4.

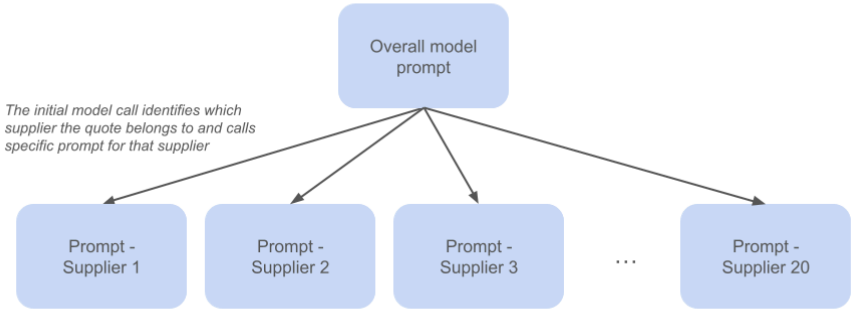

Figure 4.4 - Prompt hierarchy for refined model

To illustrate the variations between different prompts, two examples are provided below. The prompts are shortened in the section where the output is specified; in the application three fields of output are used to impose the model to generate a complete list of items, rather than restricting its response to only the initial item. The two prompts are modified to address specific differences between the suppliers and their quotations. For instance, the two suppliers are using different phrases to describe prices and quantities. For one of the suppliers, it was required to stipulate that items sharing the same item ID should be listed in separate rows. Additionally, certain information had to be specified due to missing information. For example, if an item lacked an item ID, the quotation's ID was used in its place. These adjustments were made to resolve recurring issues when interacting with this particular supplier.

- Prompt for Supplier X:
  The Quotation contains information about items that need to be extracted. The task is to extract information from the Quotation and output it as a CSV. If the information is not available, insert N/A. If there are several

items in the Quotation, please provide the information for all items and add them to the CSV output as follows
If two items have the same item id, please add the information for the second item in the next row of the CSV.Where each item is separated by ;;; and a new line and the columns are separated by ;.
The CSV output should have the format as follows where each row MUST contain 12 elements & should not be surrounded by "" & and you should not include the column names in the output:

[insert quotation number];[insert supplier];[insert item ID, if no ID available, use Quotation number];[insert item description, if no description available, use quotation number];[insert date in format: YY/MM/DD];[insert the latest date in the validity period in format YY/MM/DD];[insert original unit price per item];[insert discount, if no discount found, set to 0%];[insert unit price per item after discount, if you don't find any discount, set this to the same as original unit price per item];[insert currency in format ISO 4217 (e.g. USD, EUR, SEK etc.)];[insert quantity];[insert category based on below];;;

Category should be:
1. Spare parts if spare parts are mentioned in the quotation.
2. Installation if installation is mentioned in the item description.
3. "Material 0010" if not installation or spare parts.

“”” Extracted Text”””

- Prompt for Supplier Y
The text below marked as "Extracted text" has been extracted from a PDF that is a Quotation.
The Quotation contains information about items that need to be extracted.
The task is to extract information from the Quotation and output it as a CSV.
If the information is not available, insert N/A. If there are several items in the Quotation, please provide the information for all items and add them to the CSV output as follows
If two items have the same item id, please add the information for the second item in the next row of the CSV.
Where each item is separated by ;;; and a new line and the columns are separated by a semicolon.:

There is a data cell in the text that contains Lead Working Days, avoid this column.

The CSV output should have the format as follows where each row MUST contain 12 elements & should not be surrounded by "" & and you should not include the column names in the output:

[insert quotation number];[insert supplier: Supplier Y];[insert item No, if no item No available, use Quotation number];[insert item description, if no description available, use quotation number];[insert date in format: YY/MM/DD];[insert the latest date in the validity period in format YY/MM/DD];[insert unit price];[insert: 0%];[insert unit price];[insert currency in format ISO 4217 (e.g. USD, EUR, SEK etc.)];[insert quantity per batch];[insert category based on below];;;

Category should be:
1. *Spare parts ONLY if spare parts are mentioned in the item description.*
2. *Installation if installation is mentioned in the item description.*
3. *"Material 0010" if not installation or spare parts.*

“””Extracted text”””

The implementation of unique prompts for each supplier led to an almost perfect accuracy, suggesting the potential for achieving comparable high performance across the entire project that was previously only attainable with a singular supplier focus. This approach demonstrates the feasibility of tailoring prompts and models to enhance accuracy across diverse supplier quotations.

4.2.1.5 Evaluation

The success of the application can be measured very straightforwardly by measuring the accuracy as the number of correctly mapped fields compared to the true budget divided by the total number of fields. This was then cross-evaluated with relevant personnel at Tetra Pak to evaluate the less exact fields of the output, such as “Item Description”, where the output did not have a clearly defined ground truth. Table 4.3 gives an overview of the type of fields present in the budget used to determine the accuracy score for each iteration.

Table 4.3: Budget fields

Budget field	*Description*
Quotation number	Unique identification of quotation
Supplier name	Name of supplier

Item ID	Unique ID of a specific item in quotation
Description	Short description describing the item
Quotation date	Date of quotation received
Period of validity	Date of when quotation is no longer valid
Original Price	Price before discount
Discount	Discount for specific item
New Price	Price after discount
Currency	Currency specified in the quotation
Quantity	The number of items in each row
Exchange rate	The rate of exchange for the currency when the quotation is received

With some exceptions, each item ID corresponds to one row in the budget. For items of large quantities and low prices, items may be grouped into one row to keep the budget from being too large. For the calculation of accuracy, each item ID is considered a separate row.

4.2.2 Implementation of Demo - Risk Register Support

The process of handling the risk register also received a high WSJF score and could be a possible candidate for augmentation with AI. To initiate an AI solution, a demo was developed using the CRISP-ML(Q) framework.

4.2.2.1 Business understanding

Managing the risks of a project is a critical part of the job of a project manager (PMI, 2017). As identified in Table 4.2, a project manager at Tetra Pak spends significant time on creating a so-called risk register, containing all the risks associated with the project. In a discussion with project managers at Tetra Pak, it was determined that a chat assistant that could reason about risks in a project and make recommendations on how to mitigate risks based on past project knowledge could potentially augment the project managers' experience. As it currently stands, each project manager creates an Excel document containing all possible risks of a new project along with mitigation strategies for each risk at the start of a project. This Excel document is then updated and referenced throughout the project to support decision-making related to risks. In an ideal scenario, this would be replaced by a chat-GPT-like assistant that has access to all past projects and their corresponding risks and mitigations, which would serve as a risk assistant to the project manager. The ideal setup is outlined in Figure 4.5.

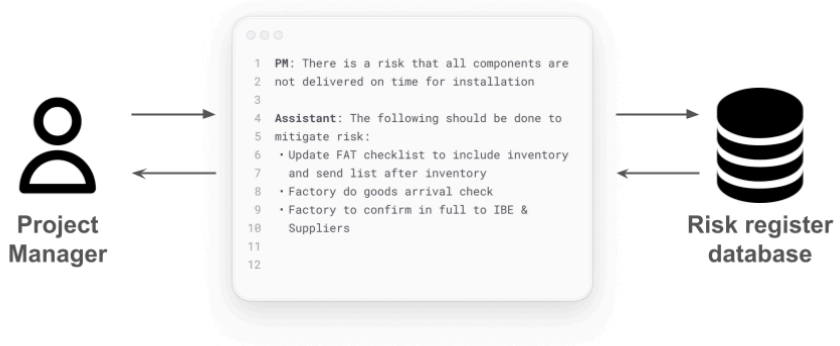

Figure 4.5: Ideal solution of risk register assistant

To test the value of an AI risk assistant, it was determined to build an assistant based on one single risk register as an early PoC.

4.2.2.2 Data understanding

The key to making an AI solution for risk register work was determined to be that the data provided is sufficient to cover the necessary risk-related knowledge. Currently, all risk registers are stored as Excel files for each project, and they are created independently by project managers. This presents a challenge, as discussed in Section 2.5.1.4, since there is no ground truth for how a risk register should be structured and what risks and mitigations are included. Another challenge is that risk registers are distributed across projects, meaning that it is not directly available and must be collected manually to be included in the model. For the initial PoC, a general risk register from which project-specific risk registers are built was used.

4.2.2.3 Data preparation

As the general risk register used for the PoC was readily available, the data preparation needed was minimal. To simplify the modeling steps, the Excel file was extracted as a CSV file.

4.2.2.4 Modeling

As described in Section 2.2.7, RAG is better suited than fine-tuning when the knowledge base referred to is updated often (Lewis et al., 2020). While the costs of fine-tuning an LLM on all available risk register data would not be unfeasible in terms of time and cost due to the limited amount of data, new risk registers would be added to the database for each project. This indicates that a RAG architecture is best suited.

The first step of the modeling phase was to create a vector database that could be accessed by the LLM when answering questions. This was done using the

LangChain library, as this is the most common and extensive tool for building RAG applications (Nvidia, 2023). First, an embedding model responsible for the vectorization of the risk data, and later the input from the user, was created using the LangChain toolbox. Then, the CSV file containing the risks and mitigations of the general risk register was embedded and vectorized to make the retrieval of information efficient. Next, the vector database and embedding model were connected to GPT-3.5 using LangChain. While a more advanced LLM like GPT-4 would likely produce better results, GPT-3.5 was used for the initial PoC due to its simplicity of implementation and low cost.

4.2.2.5 Evaluation

RAG applications can be evaluated objectively by frameworks such as RAGAS (Es et al., 2023) or subjectively by human evaluators (Tobin, 2024). Due to the simplicity of the initial PoC, it was only evaluated by human evaluators at Tetra Pak. While the PoC fulfilled the technical requirements and could successfully respond to inquiries about risks and mitigations based on the available risk register, the business potential of the solution was deemed less impactful than that of the budget PoC. It was thus determined not to continue further with the solution at this point. See Section 6 for a further discussion on why the business impact of the solution is currently limited.

5. Results and Analysis

This section presents the results obtained from the application of the methodology described in Section 4. Specifically, for the budget demonstration, the results include the accuracy scores obtained through the use of the developed software. Additionally, a business case is presented, outlining the potential of the AI solution within a practical business environment. This result with related analysis aims to assess both the technical and the economic viability of the AI solution in a real-world setting.

5.1 Budget Demo Results

The results from the first demo are presented in this section, along with recommendations for future work and a business case of expected return on investment for the solution. The results for each of the iterations of the solution are presented in Table 5.1 below.

Table 5.1: Results of demo

Model version	*Accuracy*	*Comment*
Attempt 1 - GPT-3.5 Turbo (One supplier)	50%	
Attempt 2 - GPT-3.5 Turbo (One supplier)	58%	
Attempt 3 - GPT-4 (One supplier)	94%	
Attempt 3 - GPT-3.5 Turbo - Fine-tuned (One supplier	96%	
Attempt 4 - GPT-4 Turbo (Full project) - one prompt	82%	Includes only quotations that were readable by the text extraction tool
Attempt 5 - GPT-4 Turbo (Full project) - Individual prompts	98%	Includes only quotations that were readable by the text extraction tool

There are several things worth noting about the accuracy results presented. The full project used to evaluate the final model iteration consisted of 20 suppliers with between one and five quotations each and widely varying structures of quotations. Of the 20 suppliers, two had quotations that were not readable by the text extraction tool used and are thus not included in the results. Including these would yield significantly lower accuracy scores for these models since they would all be wrong. They were excluded since it was determined that it was not reflective of the model's performance that the PDFs were not readable. When deciding if an entry was correct or not, some fields were easier than others to measure. For prices, dates, item IDs, supplier names, and quotation numbers, a correct value was determined to match exactly those of the ground truth budget. For the description field, the decision of a correct input was made based on how well the description by the model explained what the item included in terms of the product purchased and its functionality since there was no clearly defined standard for how these fields were to be filled. Due to the nature of LLMs, the result varied slightly each time the model was run. This means that the accuracy scores presented would vary slightly from time to time but generally stay within a few percentage points of the presented score.

5.1.2 Analysis of Results and Recommendations

In general, the models used performed well at the given task, even for suppliers with quotes that seemed unstructured to the human eye. It is, however, very clear that some suppliers´ quotations caused more issues than others. Specifically, quotations that were scanned PDFs were not processed at all since the text could not be extracted. Some quotations contain over 50 line items, which the model could not handle due to the output exceeding the allowed output window of the model. These quotations could possibly be handled by grouping items into one line item in the model. This was, however, not completed since it was determined to require more work than what was in scope for this report. Further, the model struggled to handle quantities of items for some suppliers, as it mixed up the quantity with other numbers in the quotation. Overall, the model would need to be further refined to perform satisfactorily in production. As discussed in Section 2, the model can generally be improved by fine-tuning and further refined prompt engineering, each of these with slightly different use cases.

Fine-tuning is generally best used when the output of the model needs a specific format (Allard & Jarvis, 2024), which is the case with this solution. Fine-tuning could be especially relevant for the supplier quotes where the model mixed up numbers in quotes, as providing a large number of training examples and desired outputs could potentially yield better results, as was seen in the fine-tuned GPT-3.5 model. For the supplier quotes with high numbers of line items, a combination of

prompt engineering and fine-tuning could potentially solve the issue. Prompting the model to calculate the total cost of all items in the quote and then fine-tuning it only to output one line containing the total could potentially solve the issue. In cases where the model cannot handle the extracted text of the PDF due to tables and other structures, a more advanced extraction method, such as using complex optical character recognition (OCR) techniques combined with machine learning techniques for structure recognition, could be used but would require high technical competence and capability (Dhouib & Bettaieb & Shabou, 2023). In cases where the model cannot even read the PDF, a discussion with suppliers is required to make the model work in this instance. In summary, the next steps of finalizing the model for production would be to:

- Consider fine-tuning specific models for suppliers with advanced structures.
- Consider a combination of prompt engineering and fine-tuning to manage quotations with large quantities of items.
- Consider implementing advanced PDF text extraction to handle complex structures in quotations.
- Consider communicating the need for non-scanned PDF documents with suppliers.

While one option for handling discrepancies in output for different suppliers would be to communicate a set template for quotations that all suppliers are told to use, this may not be a viable option due to relational reasons. The technical costs related to further improving the solution are minimal, as will be seen in the next section, and the main costs incurred are due to the man-hours related to building the solution. Due to this, it is likely more beneficial to invest in further refinement of the model to handle most use cases and simply communicate changes in quotation structures where critical. For further discussion on ways that could potentially improve the model, see Section 6.

5.1.3 Business Case

Table 5.2 presents the calculated potential payback period and potential net present value (NPV) of implementing the model using GPT-4 and a fine-tuned GPT-3.5.

Table 5.2: Budget fields

Financial measure	*Value*
Potential Total NPV (GPT-4)	€214538-556537,5
Potential Payback time (GPT-4)	0,36-0,8 years
Potential Total NPV (GPT-3.5 FT)	€224332-576788

Potential Payback time (GPT-3.5 FT)	0,34-0,78 years

The ranges include an optimistic case of full utilization of the model and a less optimistic case of 50% utilization of the model. As can be seen from both the total potential NPVs and the potential payback times, the project is expected to be highly beneficial both in the optimistic and less optimistic case. The difference between using GPT-4 and fine-tuned GPT-3.5 is minimal. The exact calculations for the optimistic cases for GPT-4 and fine-tuned GPT-3.5 are presented in Tables 5.3 and 5.4. The less optimistic cases are excluded from the thesis since they are exactly the same as the optimistic case with half the cost savings. The NPV calculation is split into four parts in both tables: yearly net present value, total yearly cash flow, positive cash flow and negative cash flow.

- **Yearly NPV:** Total yearly cash flow discounted to todays value with a 10% discount rate.
- **Total yearly cash flow:** The difference between total positive cash flow and total negative cash flow for each year.
- **Positive cash flow:** The positive cash flow is the total yearly cost savings derived from the solution. Detailed calculations for cost savings can be found in appendix B.
- **Negative cash flow:** The negative cash flow is the total costs of using the model. They include man hour costs of developing and maintaining the model, technical development costs from running and training the model in development, and runtime costs consisting of costs of API-calls in production. See appendix B for detailed calculations of all cost components.

Using a fine-tuned GPT-3.5 and GPT-4 mainly differ in the development costs being slightly higher for GPT-3.5 due to fine-tuning, as well as runtime costs being higher for GPT-4 due to higher fees for API-calls.

Table 5.3: Total costs and cost savings using GPT-4.

NPV		**2024**	**2025**	**2026**	**2027**	**2028**	**2029**
Figures in EUR	Year	0	1	2	3	4	5
NPV yearly		**−57 710,4**	**147 107,7**	**133 734,3**	**121 576,6**	**110 524**	**100 476**
Total yearly cash flow		**−57 710,4**	**161 818**	**161 818**	**161 818**	**161 818**	**161 818**
Cumulative		−57 710,4	104 108,1	265 926,5	427 745,0	589 563	751 381

investment

Positive cash flow		2024	2025	2026	2027	2028	2029
Figures in EUR	Year	0	1	2	3	4	5
Total positive cash flow		**-**	**180 000**	**180 000**	**180 000**	**180 000**	**180 000**
Total cost savings			180 000	180 000	180 000	180 000	180 000
Growth (%)				*0%*	*0%*	*0%*	*0%*

Negative cash flow		2024	2025	2026	2027	2028	2029
Figures in EUR	Year	0	1	2	3	4	5
Total negative cash flow		**57 710**	**18 182**	**18 182**	**18 182**	**18 182**	**18 182**
Man hours		57 600	11 520	11 520	11 520	11 520	11 520
Growth (%)				*0%*	*0%*	*0%*	*0%*
Development costs		110	11	11	11	11	11
Growth (%)			*−90%*	*0%*	*0%*	*0%*	*0%*
Runtime costs			6 650	6 650	6 650	6 650	6 650
Growth (%)				*0%*	*0%*	*0%*	*0%*
Other costs			-	-	-	-	-
Growth (%)				*0%*	*0%*	*0%*	*0%*

Table 5.4: Total cost and potential cost savings using GPT-3.5.

NPV		2024	2025	2026	2027	2028	2029
Figures in EUR	Year	0	1	2	3	4	5
NPV Project total		**−57 820,8**	**152 518,9**	**138 663,8**	**126 103,6**	**114 679,7**	**104 254,3**

Total yearly cash flow		**−57 820,8**	**167 770,8**	**167 783,2**	**167 843,9**	**167 902,5**	**167 902,5**
Cumulative investment		−57 820,8	109 950,0	277 733,2	445 577,1	613 479,6	781 382,1

Positive cash flow		**2024**	**2025**	**2026**	**2027**	**2028**	**2029**
Figures in EUR	Year	0	1	2	3	4	5
Total positive cash flow		**-**	**180 000**	**180 000**	**180 000**	**180 000**	**180 000**
Total cost savings			180 000	180 000	180 000	180 000	180 000
Growth (%)				*0%*	*0%*	*0%*	*0%*

Negative cash flow		**2024**	**2025**	**2026**	**2027**	**2028**	**2029**
Figures in EUR	Year	0	1	2	3	4	5
Total negative cash flow		**57 821**	**12 229**	**12 217**	**12 156**	**12 097**	**12 097**
Man hours		57 600	11 520	11 520	11 520	11 520	11 520
Growth (%)				*0%*	*0%*	*0%*	*0%*
Development costs		221	44	45	50	50	50
Growth (%)			*−80%*	*2%*	*10%*	*0%*	*0%*
Runtime costs			665	652	587	528	528
Growth (%)				*−2%*	*−10%*	*−10%*	*0%*
Other costs			-	-	-	-	-
Growth (%)				*0%*	*0%*	*0%*	*0%*

6. Discussion

This chapter discusses the theoretical frameworks, methodologies, and results derived from this thesis. It addresses the limitations and trade-offs associated with the employed methodologies, providing critical reflections on these aspects. Additionally, the chapter evaluates the research outcomes, considering what improvements could be implemented to the derived demos. Lastly, it outlines general challenges encountered during the research process and suggests directions for future research.

6.1 Limitations of the used Framework

As mentioned by Ångström et al. (2023), introducing small pilot projects for AI solutions is a key step in introducing the use of AI in organizations. It creates awareness and understanding of the potential of such solutions among employees. Building on this foundation, the framework for identifying tasks with AI potential introduced by Brynjolfsson & Mitchell (2017) combined with the WSJF approach by Reinertsen (2009) is a good starting point for identifying and prioritizing the pilot projects that are likely to have the biggest impact on the organization. However, as organizations move past the initial stages of AI discovery, the approach has potential drawbacks. While the approach may foster incremental process innovation in organizations, it is unlikely to lead to disruptive innovation (Satell, 2017). It is not uncommon for organizations to introduce entirely new business models related to AI, something that is unlikely to be the case when applying this approach (Mariani, Machado & Nambisan, 2023).

Furthermore, the framework is applied exclusively to the work of one project manager at Tetra Pak. While the identified tasks have been discussed with several

stakeholders and other project managers to give a wider perspective, the tasks themselves and their priority presented here are based almost exclusively on the opinions of one project manager. This inevitably introduces bias to the tasks being selected, both initially as they are identified and as they are weighted and scored for prioritization (Ferrera, 2023a). The decision to include only one project manager in the selection process was based on time constraints, as well as the assumption that his long experience in the company and large number of completed projects would provide a full view of the tasks of a project manager at Tetra Pak. However, the framework would ideally be applied to the work of all project managers in the division and the tasks identified based on the opinions of the full team. Furthermore, there might be tasks performed by other project team members that have a high potential for AI use that are not covered here. Ideally, the framework would be applied to all project team members, as well as to the portfolio management team, to identify all aspects of the project process that can be augmented or replaced by AI.

Finally, the four aspects used to decide if a task has basic AI potential are based on the framework by Brynjolfsson & Mitchell (2017), as well as how GenAI differs from traditional machine learning (Weisz et al., 2024). Based on these two sources, the four aspects are valid criteria to exclude tasks from further work. The final decision of what tasks to exclude was however somewhat subjective and based on discussion between the project manager and the authors of this thesis. In the future, these criteria are likely going to change as AI becomes more capable of completing complex tasks and can be trusted to handle more critical aspects of the project.

6.2 Limitations of the Budget Demo

A discussion of the results of the budget demo and possible improvements to the solution are presented in this section. The demo highlights challenges that could potentially be addressed by future technological advancements yet currently need to be considered to enable the model to be fully production-ready. Improvements are related to introducing complex workflows related to error handling, individually fine-tuned models, implementation of tools, and improved data extraction from PDF files.

6.2.1 Choosing the right Foundation Model

In the budget demo developed for this thesis, two versions of OpenAI's models were utilized: the older GPT-3.5 and the newer GPT-4. When used in the application, their performance across different sets of prompts varied significantly,

from 50% using GPT-3.5 and a simple prompt to 98% using GPT-4 and individual prompts for each supplier. However, fine-tuning GPT-3.5 to a specific supplier resulted in 94% accuracy and costs only one-tenth to use compared to GPT-4 (OpenAI, 2024). This variation underscores the importance of understanding model-specific capabilities and limitations in different contexts. Fine-tuning requires significant effort, and model choices are a weigh-off between the effort required to build the model and the costs of running it. As seen in the financial model of the application, the personnel costs of building the application are much more significant than the technical costs, weighing in favor of using the more capable GPT-4 and spending less time on building bespoke models using GPT-3.5. The technical enhancements from GPT-3.5 to GPT-4 include more sophisticated training algorithms and a larger dataset, which likely contributed to the observed performance variation between the models (OpenAI, 2023c).

Despite these advancements, GPT-4 exhibits inherent challenges, such as output biases and high computational demands, which could hinder increased accuracy. The anticipated development of more advanced AI models will most certainly address many of the current limitations, potentially leading to a more accurate solution in this application of AI. For example, new GPT models are expected to be launched already this year with new and better capabilities (Hays & Rafieyan, 2024).

Another limitation observed in the study involved the handling of certain quotations that triggered excessively long responses, occasionally leading to application crashes. Today, GPT-4 has a limited output response to a maximum number of tokens (OpenAI, 2024b). For this specific case it means that it cannot handle quotations with a large number of line items. To solve this problem in the present, a potential workaround is to manually adjust the model to generate shorter responses that do not encompass the entire quotation and instead summarize line items into one. However, it is likely that future iterations of GPT models will address this challenge by allowing longer output windows similar to the latest Gemini models (Pichai, 2024).

6.2.2 Handling Errors

The inherent uncertainty in the outputs of LLMs requires the implementation of robust safety mechanisms to alert users when potentially erroneous conclusions are drawn. This can be achieved through multiple strategies addressed by the People + AI Guidebook (n.d). Firstly, rigorous error-handling protocols could be established using traditional means. For instance, data validation rules could ensure that inputs conform to expected formats and ranges; valid date formats, positive integers for quantities, etc. Additionally, proactive warning systems could be implemented, where, for example, quantities exceeding a high number trigger

alerts for manual review, enhancing oversight and minimizing the risk of inserting wrong data into the budget.

Another potential approach involves the use of one LLM to evaluate the output of another. This method, leveraging the concept of inter-model evaluation, may be particularly valuable for increasing accuracy and reliability in AI solutions (Arthur team, 2023). By establishing a continuous feedback loop, where a pre-trained or generic LLM reviews and assesses outputs, the process could not only automate error detection but could also significantly reduce the need for human intervention, thereby increasing the security and robustness of the application. A potential example of this could be to let the model run several times, saving the output for each run. A separate LLM could then evaluate the outputs and identify where specific fields differ between runs, indicating errors. These could then be cross-checked again by the LLM, providing a more robust model.

6.2.3 Fine-tuning models for each supplier

A potential enhancement to the demo could involve further specialization tailored to each supplier's specific needs. Currently, individualized prompts are designed to capture unique information relevant to each supplier, such as names or structural data. In the initial PoC for the budget application, where only quotations from one supplier were utilized, fine-tuning was employed not only to enhance the accuracy of the model but also to improve its efficiency and cost-effectiveness (OpenAI, 2023d). Given that a fine-tuned model builds upon a pre-trained base, it inherently processes queries more rapidly, and the possibility of using a more cost-effective model such as GPT-3.5 is possible (Shin et al., 2023; Fatemi & Hu, 2023). Expanding this approach and implementing a fine-tuning strategy for each supplier could optimize the model’s performance relative to the specific needs of each distinct structure and layout associated with each supplier's quotation.

An approach that could potentially improve model accuracy and decrease initial development spend is to fine-tune the AI model continuously as a part of the application's operational cycle (Microsoft, 2024). This process would involve real-time incorporation of new data obtained from manually processed quotations directly into the model's training dataset. When quotations are entered and successfully validated without errors, they could automatically contribute to the model’s ongoing learning. If the model outputs an erroneous field, the project manager could correct this field which would then serve as further training data for the model. Such a mechanism would establish a positive feedback loop whereby the accumulation of validated data incrementally enhances the model's performance over time, mimicking the approach of reinforcement learning by human preference used to train chat-bots (Christiano, 2023). This adaptive

learning approach not only refines the model's accuracy but also ensures that it evolves in alignment with emerging data trends from the quotations.

6.2.4 Implementing tools

Another potential enhancement to the current implementation could involve the integration of tools within the AI solution. According to LangChain (2024), these tools enable the application to interact more effectively with its environment. One practical use case could involve the AI system interacting directly with the file system that stores all quotations. This would allow the AI to effectively monitor and distinguish between relevant and outdated quotations, enhancing the efficiency of data management. Furthermore, another tool could be developed to manage and summarize information from extensive lists of items within the quotations. This tool would enable the application to condense large volumes of data into more manageable and relevant summaries, thus improving the accessibility and usability of the information. There is currently not a lot of finished research on the topic, but it is likely to be an interesting topic for future research (Ng, 2024b).

6.2.5 Improving Data Extraction

As described in Section 4.3.1.3, a challenge encountered in the application relates to the extraction of data from PDF documents. Ideally, if the quotation data were delivered to Tetra Pak in a structured format, a straightforward automation script could facilitate the extraction of information and its subsequent insertion into Excel. While the data in a PDF document can be easily interpreted by humans, the inherent structure proves challenging for a computer. The current software, PDFminer, manages to extract text from the majority of quotations in the format depicted in Appendix A. This text, in comparison with the PDF document, is feasible to process through an AI model.

Some of the PDFs are scanned documents that do not contain information in the source code of the PDF. To be able to read these types of documents, other techniques are needed, such as Optical Character Recognition that converts images into machine-encoded text that the computer can read (Dhouib & Bettaieb & Shabou, 2023). For the current method, the extracted information lacks structure. As noted by Zaman et al. (2020), a more intricate extraction process could impose more structure. It is anticipated that the budget application would perform more effectively in mapping the extracted values to their corresponding columns if the information used was more structured.

6.3 Challenges with Implementing the Model in Practice

The AI solution is provided with inherent capabilities that may lead to challenges both in its development and operation. As mentioned in Section 2.5.1.2 and 2.5.1.3, an AI model is susceptible to specific challenges regarding bias and hallucinations. Additional challenges relate to organizational and cultural aspects. Central to these challenges is the necessity of cultivating a data-driven culture to foster trust in the AI solution, coupled with the implementation of comprehensive educational programs to ensure that personnel utilize the systems effectively.

6.3.1 Bias and Hallucinations

Data-driven bias is a consequence of the training data, which may contain inherent biases (Dwivedi et al., 2023). Such biases are subsequently replicated by the model during its operation. In the context of this application, the training data needs to be in the same structure as new unseen data to ensure replication of expected behavior. However, if the structure of the input data alters due to changes such as a modification in the quotation format from a supplier, the model, trained on the previous data structure, may fail to recognize or adapt to the new format. This issue is compounded by automation bias, the tendency of humans to accept computer-generated outputs as accurate, stemming from an overreliance on technological solutions (Mosier & Skitka, 1998). Such bias may lead to operational failures if users of the AI system place trust in its outputs without sufficient scrutiny (Potaznik, 2023). Moreover, the phenomenon of hallucinations in AI models exacerbates these issues (Ji et al., 2023). In this context, hallucinations refer to the generation of nonsensical or irrelevant content by the AI model in an attempt to provide a solution, particularly when the system fails to locate the correct data (e.g., item quantities in quotations) and instead resorts to making baseless predictions (Fui-Hoon Nah et al., 2023). The interplay among data-driven bias, automation bias, and hallucinations can create a cascading effect, where one issue gives rise to another, potentially amplifying minor inaccuracies into significant system failures.

To mitigate these challenges, Ferrera (2023b) advocates for the implementation of human-in-the-loop approaches, which integrate human oversight at critical stages of the AI operation. This perspective aligns with the strategies proposed by Park et al. (2019), who emphasize the importance of manual analysis of AI-generated solutions before their acceptance. Additionally, Gill and Kaur (2023) support the notion of continuous monitoring and the adoption of adaptive systems that can adjust to evolving data structures and operational contexts. These methodologies aim to enhance the reliability and accuracy of AI systems by ensuring they remain under human supervision and are capable of adaptation in response to dynamic

environments. For the application at hand, human-in-the-loop approaches can include safety systems, explained by People + AI Guidebook (n.d), that force the user to go through results that are outside certain thresholds manually. This ensures that any anomalies detected by the AI in the extraction process, such as unusually high quotation amounts or mismatched product categories, are flagged for human review before being entered into the budget.

This approach could be used in liaison with continuous prompt engineering and fine-tuning of the AI model, which could align the systems with changes in the structure or content of the PDFs. As suppliers might frequently update their document formats, the AI system could be designed to learn from these changes adaptively. For example, if a supplier shifts from listing items vertically to a horizontal format, the AI should be capable of recognizing this shift and adjusting its data extraction logic accordingly. Both the fine-tuning and the prompt engineering need human intervention in the form of locating the changes and adding them to the dataset or prompt window. For minor adjustments in structure, continuous fine-tuning may directly resolve issues, but for more substantial alterations, the system may require both fine-tuning and modifications to the prompts as described by Allard & Jarvis (2023).

6.3.2 Lack of Trust

The converse implication of implementing AI solutions is the potential development of distrust in both the derived data and the AI systems themselves. As discussed in Section 2.5.3.2, an application such as the one proposed in this thesis may encounter skepticism stemming from a lack of explainability in the results it produces (Gill & Kaur, 2023). This could, in combination with a general worry towards AI and the fear of job displacement, lead to a fear of utilizing the solution (Arslan et al., 2021). Should an application remain underutilized, the value of the solution decreases significantly, regardless of its operational performance. To prevent such outcomes, a cultural shift toward embracing data-driven solutions is imperative. Ångström et al. (2023) highlight that there are often misaligned expectations regarding AI, where stakeholders may not fully understand or appreciate the capabilities and limitations of AI technologies. Addressing these challenges involves the adoption of strategies to cultivate an organizational culture that not only trusts but is also proficient in data-driven technologies. This requires transparent communication regarding the functionality and decision-making processes inherent in AI systems. Additionally, it involves implementing robust training programs aimed at enhancing employee comfort with and proficiency in these technologies.

Through education, the expectations surrounding the AI solution can be accurately calibrated, preventing the formation of overly ambitious expectations that might amplify issues such as data-driven bias, automation bias, and hallucinations (Ångström et al., 2023; Shrivastav, 2023; Beauchene et al., 2023). Conversely, it also avoids setting expectations too low, which could lead to underutilization of the solution. Nilsson et al. (2024) highlight that while project managers have high expectations for how AI will impact project management, only 35% possess basic knowledge of AI technologies. This disparity underscores the need for educational programs among project managers to ensure that expectations are both realistic and informed.

6.3.3 Data Protection

In Section 2.5.1.1 the challenge of safeguarding data within AI applications is underscored. The application involves processing of information to third-party models such as OpenAI's GPT-3 and GPT-4. As highlighted by Fui-Hoon Nah et al. (2023) and Gill & Kaur (2023), this raises concerns regarding the potential disclosure of sensitive information. In the context of the prototypes in this thesis, the data processed includes personal names, addresses, and detailed information concerning products and suppliers. To facilitate the development of the prototype while mitigating privacy risks, all sensitive information was anonymized using a Python script. This script replaced all personal and potentially classified data with fictitious, non-sensitive data. This anonymization process did not alter the structure of the documents, thereby preserving the validity of the experimental results.

The data protection issues become more pronounced and complex when considering the transition from a prototype to full production. Anonymizing data for daily operational use would be impractical and time-consuming, potentially diminishing the application's value. Several strategies are available to address these concerns. One option involves relying on the OpenAI API's security measures. According to OpenAI (2024b), the organization assures users that it does not share user data, maintaining that enterprises retain complete ownership and control over their data. Alternatively, as advocated by Hitaj (2017), deploying locally trained AI models could significantly reduce the risk of external data exposure. While local models offer enhanced data security, they may lack the advanced capabilities and performance benefits of larger, more sophisticated models provided by services like OpenAI. The choice between using third-party APIs and a local AI solution involves a trade-off between operational efficiency and the sensitivity of the data involved.

6.3.4 Making Changes to the Process versus the Technology

As mentioned in Section 5, there were some supplier quotations in the budget demo that were not readable by the model. In this case, it is necessary for the supplier quotation to be updated so that the model works properly in production. Taking this a step further, the model would have likely been unnecessary in the first place if the supplier quotations had been standardized. If they were, a simple Python script could have extracted the information and input it into the Excel file without the need to use an LLM to transcribe the PDFs. This builds on the discussion around the suitability of using AI for a task by Brynjolfsson & Mitchell (2017). Even though a task may be technically suitable to augment with AI, it may not be the best possible option to solve the problem at hand. In this specific case, changing how suppliers do quotations would require significant effort and would likely be harder than building an AI model. In other use cases, the right solution may well be to change the process instead. This highlights the importance of establishing a business understanding before developing a model and why it is important to include both data experts and domain experts in the initial discussions (Studer et al., 2021).

6.3.5 Measuring Value Creation and Return on Investment

The budget demo will not generate any new revenue or direct cost benefits, but will instead save time for project managers and thus result in potential cost savings. This is a common issue when measuring the return on investment for AI solutions, as the potential cost savings are dependent on the utilization of the model in the organization (Enholm et al., 2022). For this specific case, the potential cost savings will only be realized if project managers are able to spend the time gained on more value-creating activities within the organization or if the usage of the model requires fewer project managers. While uncertainty around future cash flows may create a reluctance to invest in new technology, it is key to consider the different potential scenarios when evaluating investment opportunities (Enholm et al., 2022).

Commonly, the base scenario against which potential investments are measured (the base case scenario) is “business as usual,” i.e., doing nothing. However, for new technology, “business as usual” may result in negative future cash flows if competitors implement the new technology (Koller, Goedhart, Wessel, 2020). In this case, Tetra Pak may find it hard to attract talented project managers in the future if business is kept as usual and competitors are automating repetitive tasks such as budget creation with the help of AI. This is an example of why it is generally a good idea to consider aspects other than pure monetary benefits in the organization when evaluating future investments in AI (Borges et al., 2021).

6.3.6 Responsibility for Model Results

As mentioned by Solaiman (2023), the person who is held responsible for the results of an AI model is often unclear in organizations. As discussed in several instances in this report, the budget model has the potential to cause erroneous results due to issues related to LLMs, such as hallucinations and bias. It is, therefore, critical to assign responsibility for the results of the model within Tetra Pak. The most reasonable way to assign responsibility is to put the project manager, who is responsible for the input in the budget, to be responsible for the input values to the budget generated by the model. However, it is also critical to assign responsibility for the model itself to a technical expert who can troubleshoot issues if the budget application starts to produce erroneous results regularly. This is supported by Studer et al., (2021), who highlight that machine learning models need constant maintenance to ensure that they keep producing satisfactory results after deployment.

6.4 General Challenges

In the process of prioritizing possible AI solutions, several tasks with high potential were assigned lower scores, influenced by factors other than their project impact. A primary issue identified was related to data management inefficiencies and limitations. As detailed in Section 4.1, many tasks that could have been addressed were hindered by data-related dimensions such as data availability. For instance, in predicting contingency, a barrier was the lack of data availability. This is a common cause of failure in implementing AI in organizations, as discussed by Ångström et al. (2023). In this case, the data was not centralized in a data storage, but rather dispersed across factories globally. This meant that the data was not accessible centrally for use in AI models, and also lacked clear standards for how it was created in the first place. As discussed by Tran (2023), the key to solving these issues is to have data easily accessible to the consumer of data and set standards for what data is stored and how.

Another challenge related to data management that hindered the utilization of AI was encountered in the prediction of travel costs. Although the necessary data was available in centralized data storage, there were several challenges in extracting the relevant data due to inadequate knowledge about the data and the systems in which it was stored. This scenario exemplifies what Ångström et al. (2023) describe as a collaboration issue. Effective collaboration between data scientists and operational experts could have potentially resolved these issues. To enhance the development and implementation of AI solutions, adopting an integrated approach where data scientists work alongside project managers to address these challenges could prove highly beneficial.

Following the initial prioritization, the two PoCs were developed. As noted in Section 4.2.2, the risk register solution with a RAG application was not pursued. The demonstration of this solution revealed several drawbacks. It was concluded that the RAG solution provided information that, while accurate, was accessible and well-known knowledge that didn't provide real value for the project manager who used the application. Simply answering questions about risks and mitigations did not provide substantial value to experienced project managers. In this instance, the data within the risk register alone was not inherently valuable but could have been in combination with other data that could have provided unique insights. This issue is highlighted by Bensen (2024). A frequent challenge in real-world RAG implementations is the quality of the underlying dataset. To derive substantial value from such applications, the data must be structured, cataloged, and accessible. Unfortunately, the solution in this instance does not possess a dataset with these attributes. A more effective strategy would have been to merge the risk register with data from lessons learned, allowing for the anticipation of risks based on previously completed projects. This approach could have provided significant value; however, it was not feasible due to the absence of adequate lessons learned data.

As described in Section 2.5.4, the development of LLMs from scratch poses significant environmental impacts. These environmental concerns must be addressed in scenarios where a company like Tetra Pak might consider developing its own in-house models. When utilizing an already trained model, opinions on the environmental impact vary. Kaspersen (2023) suggests that making API calls to an already-trained LLM results in minimal environmental impact. Naughton (2023) argues that while a small number of inferences with the model might not significantly affect the environment, the cumulative effect of many users may lead to a substantial environmental impact. However, there are cases where the environmental impact may be decreased by the usage of LLMs. For instance, Tomlinson et al. (2024) provide a perspective that AI systems emit less carbon dioxide per page of text generated compared to traditional human writers. Similarly, AI-based illustration systems are reported to emit less than their human counterparts. These findings suggest that in certain applications, such as writing and illustrating, the use of AI can indeed be a more environmentally sustainable option than traditional methods. These conflicting viewpoints highlight the complexity of evaluating the environmental impact of AI technologies.

6.5 Future Research Areas

This thesis has highlighted the ways in which AI can be applied in a project management setting. Specifically, at Tetra Pak, there are several areas of future research that are worth pursuing. First, as mentioned earlier in this discussion, there were several tasks identified with a high potential to be automated or augmented by AI that do not currently have the prerequisite conditions to be applied in the organization. Spending more time to create a technical and organizational environment where these solutions could be implemented is an interesting topic for future research within the company. It would likely include redesigning parts of the current data management process. Another area worth exploring is applying the framework designed in this thesis to all areas of the CAPEX implementation department to give a full picture of where AI has the highest potential in the organization. For example, managing the full portfolio of projects is likely more data-driven than running a single project which could potentially provide enough data to run more advanced models. Finally, continuing the work on the budget demo built in this thesis is a valid future research topic in itself, as building a fully production-ready model that handles all types of projects and supplier quotations would require significant effort and skill.

In general terms, the field of applying AI in different business scenarios is one of the hottest research topics today, with breakthroughs happening almost daily. Several of the issues related to the demo model in this thesis may be solved automatically with the next generation of LLMs and, thus not need any additional effort. However, there are areas of research that could be of specific interest to apply in a project management setting. Agentic workflows, where several LLMs work together in longer workflows to achieve complex results, is a research topic with the potential to have a large impact on organizations (Ng, 2024). With agentic workflows, the agent could potentially not only reply to answers and make suggestions of actions but actually interact with other software in long-running workflows that require high levels of logical reasoning. For example, Cognition Labs recently launched an LLM-based agent called “Devin,” which is stated to be able to act as a software developer that can build software from scratch and put it into a production environment on its own (Cognition Labs, 2024). Further research on how long-running agentic workflows can be incorporated to support project management processes is an interesting topic with a high potential for impact.

7. Conclusion

This section concludes the thesis, summarizes its scientific contributions, and answers the research questions.

There are nearly unlimited ways of implementing AI in project management. The most common use cases are mainly related to how individual project managers can use ready-made tools to take meeting notes, automate email responses, help with idea generation, or reason about project risks. These use cases will likely not revolutionize the way large organizations like Tetra Pak run their projects, but they may help individuals become more efficient. It is thus worthwhile for project managers to stay up to date with the latest tools to make informed decisions about what tools are most relevant to their way of working. For larger-purpose AI applications, a more structured approach is recommended. This thesis set out to answer the following research questions related to implementing AI in a project-oriented organization:

- How can teams identify and prioritize areas of the project management process that have the potential to be augmented by AI?
- How can AI be implemented in the identified areas to maximize project impact?
- What prerequisites are critical to ensure the success of AI implementation?

To answer the first question, this thesis builds on the work of Brynjolfsson & Mitchell (2017) to develop a structured approach that first identifies the tasks of the different roles in the team that have the potential to be automated or augmented by AI. The original framework is adjusted to modern GenAI use cases and applied to the context of project management. This thesis then extends upon the work of Reinertsen (2009), adjusting the original Weighted Shortest Job First framework to AI use cases. The adjusted framework is used to prioritize the tasks that have the highest impact on projects from a financial, time, and risk perspective and weigh these factors against the effort required to implement the solution for the given task.

For the second question, this thesis applies the CRISP-ML(Q) framework developed by Studer et al., (2021) to implement the solutions prioritized in the previous step. This section should be seen as a direct contribution to how AI can be applied in project management. It highlights the collaboration between

technical experts and domain experts to establish a strong understanding of both the business and technical aspects required for the solution. It also highlights the need to iterate quickly to gather feedback and improve the solution continuously, starting with developing a pilot demo that can be reviewed by key stakeholders.

Finally, this thesis brings up a number of prerequisites for the successful implementation of AI in organizations. These are based on current available research in the field and cover technical, organizational, and cultural aspects. On a technical level, it is highlighted that the data management processes of the organization need to ensure that data is easily available and of high enough quality for the AI solutions to work properly. A discussion around the need to enforce policies on data protection and mitigations against bias and hallucination of models is also brought up. On an organizational level, the thesis discusses the need for education on the potential and shortcomings of AI among project managers. Leaders also need to consider how they evaluate the value creation of AI solutions, and responsibility for the results of the solution needs to be communicated clearly. Lastly, as the value created by AI is closely tied to how well it is utilized in the organization, it is critical to establish a culture of trust in the models while maintaining a healthy dose of critical thinking. The contribution of the thesis on this subject is the exemplification of how these challenges were present in the processes of implementing AI at Tetra Pak.

Based on what is discussed in this thesis and previous research, it seems very likely that AI will continue to be central to the business processes of companies in the future (Chen et al., 2023). This has led to high enthusiasm towards applying AI to a large number of use cases (Nilsson et al., 2024; Ångström et al., 2023). However, AI is not the cure for all problems. The publicly available AI solutions are often not adaptable to company use cases and miss the mark when measuring the value created (Gartner, 2019). This is both due to the AI solutions themselves since the technology is still maturing, and that the range of potential business applications is still limited. Companies are instead forced to hire data scientists to build bespoke solutions, something that often also fails due to unreasonable expectations and insufficient communication between domain experts and data science teams (Ångström et al., 2023). This is not to mention the challenges related to privacy and security that come with increased AI use, both from a general public perspective and from an organizational perspective (Solaiman, 2023; Gill & Kaur, 2023). Matters related to AI security and governance will undoubtedly become increasingly central in proportion to the impact and usage of AI.

To take advantage of the potential of AI and not risk falling behind competitors, companies need to aggressively upskill the workforce, either through external partnerships or through internal education. This will also reflect on academia, as the demand for graduates with knowledge about these topics will increase. This is

already reflected in the hiring patterns of large corporations and is likely to keep increasing over time (Amazon Web Services, 2023). If companies are able to adapt to the change, it seems reasonable that the future will bring a combination of improved AI technology with more use cases and a lower barrier to value, together with increased knowledge among organizations. This outlook is very promising for the future of AI among organizations.

References

Asadi, A., Alsubaey, M., & Makatsoris, C. (2015). “A machine learning approach for predicting delays in construction logistics”. International Journal of Advanced Logistics. Informa UK Limited. doi:10.1080/2287108x.2015.1059920.

Allard, J., & Jarvis, C. (2024). “A Survey of Techniques for Maximizing LLM Performance”. OpenAI. Available at: https://www.youtube.com/watch?v=ahnGLM-RC1Y (Accessed 4 March 2024)

ANSI (2024). “Trade Associations: Project Management Institute”. https://www.standardsportal.org/usa_en/trade_associations/pmi.aspx. (Accessed: 19 January 2024)

Arslan, A., Cooper, C., Khan, Z., Golgeci, I. & Ali, I. (2021). “Artificial intelligence and human workers interaction at team level: a conceptual assessment of the challenges and potential HRM strategies”, International Journal of Manpower. Emerald. doi:10.1108/ijm-01-2021-0052

Arthur Team. (2023) “LLM-Guided Evaluation: Using LLMs to Evaluate LLMs”, 29 September 2023, Available at: https://www.arthur.ai/blog/llm-guided-evaluation-using-llms-to-evaluate-llms (Accessed 16 April 2024)

Atkinson, R. (2016). “Technology may disrupt occupations, but it won’t kill jobs”. Monthly Labor Review. Bureau of Labor Statistics. Available at: https://doi.org/10.21916/mlr.2016.8.

Amazon Web Services and Access Partnership (2023) Accelerating AI Skills: Preparing the Workforce for Jobs of the Future. November. Available at: https://www.aboutamazon.com/news/aws/how-ai-changes-workplaces-aws-report (Accessed: 19 March 2024)

Autor, D., Levy, F. & Murnane, R. J. (2003). “The skill content of recent technological change: An empirical exploration”. The Quarterly Journal of Economics 118(4). 1279–1333.

Bansal, V. (2023). “Project Management: Planning and Scheduling Techniques”. 1st ed. Routledge. https://doi-org.ludwig.lub.lu.se/10.1201/9781003428992

Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021) “On the Dangers of Stochastic Parrots,” Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT ’21: 2021 ACM Conference on Fairness, Accountability, and Transparency, ACM. https://doi.org/10.1145/3442188.3445922

Bensen, S. (2024). “The Practical Limitations and Advantages of Retrieval Augmented Generation (RAG)”. Towards Data Science. 15 April 2024. Available at: https://towardsdatascience.com/the-limitations-and-advantages-of-retrieval-augmented-generation-rag-9ec9b4ae3729 (Accessed 26 April 2024)

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J.Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D.E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P.W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X.L., Li, X., Ma, T., Malik, A., Manning, C.D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J.C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J.S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A.W., Tramèr, F., Wang, R.E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S.M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K. and Liang, P. (2021). “On the Opportunities and Risks of Foundation Models”. arXiv. https://doi.org/10.48550/ARXIV.2108.07258.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D. (2020) “Language Models are Few-Shot Learners.” arXiv. https://doi.org/10.48550/ARXIV.2005.14165

Berntsson Svensson, R. and Taghavianfar, M. (2020) “Toward Becoming a Data-Driven Organization: Challenges and Benefits,” Research Challenges in Information Science. Springer International Publishing. doi.org/10.1007/978-3-030-50316-1_1

Brakemeier, H., Gerbert, P., Hartmann, P., Liebl, A., Schamberger, M. & Waldmann, A., 2024. Applying AI: How to find and prioritize AI use cases. Garching: UnternehmerTUM GmbH. Available at: https://aai.frb.io/assets/files/AppliedAI_Whitepaper_UseCase_Webansicht.pdf (Accessed 24 April 2024)

Brynjolfsson, E. and Mitchell, T. (2017) “What can machine learning do? Workforce implications”. Science. American Association for the Advancement of Science (AAAS). https://doi.org/10.1126/science.aap8062

Brynjolfsson, E., Mitchell, T., & Rock, D. (2018). “What Can Machines Learn, and What Does It Mean for Occupations and the Economy?”. AEA Papers and Proceedings, 108, 43–47. https://doi.org/10.1257/PANDP.20181019

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M.T. and Zhang, Y. (2023). “Sparks of Artificial General Intelligence: Early experiments with GPT-4”. arXiv. https://doi.org/10.48550/ARXIV.2303.12712

Beauchene, V., Bedard, J., Jefson, J. and Vaduganathan, N. (2023) 'How to Attract, Develop, and Retain AI Talent', Boston Consulting Group, 16 May. Available at: https://www.bcg.com/publications/2023/how-to-attract-develop-retain-ai-talent (Accessed 19 March 2024)

Borges, A.F.S., Laurindo, F.J.B., Spínola, M.M., Gonçalves, R.F., Mattos, C.A., (2021). “The strategic use of artificial intelligence in the digital era: Systematic literature review and future research directions”. International Journal of Information Management. doi.org/10.1016/j.ijinfomgt.2020.102225.

Blackman, R. and Vasiliu-Feltes, I., (2024). “The EU’s AI Act and How Companies Can Achieve Compliance”. Harvard Business Review, [online] Available at: https://hbr.org/2024/02/the-eus-ai-act-and-how-companies-can-achieve-compliance?registration=success (Accessed 21 March 2024).

Budhiraja, S. S. (2018). Extracting Specific Text From Documents Using Machine Learning Algorithms. Thesis of computer science, Lakehead University, Canada.

Campbell, M., Hoane, A.J., Jr. and Hsu, F. (2002) .*Deep Blue*. Artificial Intelligence. Elsevier BV. https://doi.org/10.1016/s0004-3702(01)00129-1.

CEDPO (2023). “Generative AI: The Data Protection Implications”. CEDPO AI Working Group. 16 October 2023. https://cedpo.eu/wp-content/uploads/generative-ai-the-data-protection-implications-16-10-2023.pdf (Accessed 18 March 2024)

Christiano, P.F., Leike, J., Brown, T.B., Martic, M., Legg, S. & Amodei, D., 2023. Deep Reinforcement Learning from Human Preferences. arXiv preprint arXiv:1706.03741v4.

Chang, Yupeng, Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Yi, Yu, P.S., Yang, Q. and Xie, X. (2024) “A Survey on Evaluation of Large Language Models”. ACM Transactions on Intelligent Systems and Technology. Association for Computing Machinery (ACM). https://doi.org/10.1145/3641289.

Chui, M., Yee, L., Hall, B., Singla, A. & Sukharevsky, A. (2023). “The state of AI in 2023: Generative AI’s breakout year”. McKinsey & Company. Available at: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year#/ (Accessed 28 February 2024)

Cichy, C. and Rass, S. (2019). “An Overview of Data Quality Frameworks”. IEEE Access. Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/access.2019.2899751.

Cognition Labs, 2024. Introducing Devin. [online] Available at: https://www.cognition-labs.com/introducing-devin (Accessed 18 Apr 2024).

Dhouib, M., Bettaieb, G. and Shabou, A. (2023) “DocParser: End-to-end OCR-Free Information Extraction from Visually Rich Documents,” Lecture Notes in Computer Science. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-41734-4_10.

Dzhusupova, R., Bosch, J. and Olsson, H.H. (2024). “Choosing the right path for AI integration in engineering companies”. *A strategic guide.* Journal of Systems and Software. Elsevier BV. https://doi.org/10.1016/j.jss.2023.111945.

Dwivedi, Y.K., Kshetri, N., Hughes, L., Slade, E.L., Jeyaraj, A., Kar, A.K., Baabdullah, A.M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M.A., Al-Busaidi, A.S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., Carter, L., Chowdhury, S., Crick, T., Cunningham, S.W., Davies, G.H., Davison, R.M., Dé, R., Dennehy, D., Duan, Y., Dubey, R., Dwivedi, R., Edwards, J.S., Flavián, C., Gauld, R., Grover, V., Hu, M.-C., Janssen, M., Jones, P., Junglas, I., Khorana, S., Kraus, S., Larsen, K.R., Latreille, P., Laumer, S., Malik, F.T., Mardani, A., Mariani, M., Mithas, S., Mogaji, E., Nord, J.H., O’Connor, S., Okumus, F., Pagani, M., Pandey, N., Papagiannidis, S., Pappas, I.O., Pathak, N., Pries-Heje, J., Raman, R., Rana, N.P., Rehm, S.-V., Ribeiro-Navarrete, S., Richter, A., Rowe, F., Sarker, S., Stahl, B.C., Tiwari, M.K., van der Aalst, W., Venkatesh, V., Viglia, G., Wade, M., Walton, P., Wirtz, J. and Wright, R. (2023). “Opinion Paper: ‘So what if ChatGPT wrote it?’ Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy”. International Journal of Information Management. Elsevier BV. https://doi.org/10.1016/j.ijinfomgt.2023.102642.

Dwivedi, Y.K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P.V., Janssen, M., Jones, P., Kar, A.K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., Medaglia, R., Le Meunier-FitzHugh, K., Le Meunier-FitzHugh, L.C., Misra, S., Mogaji, E., Sharma, S.K., Singh, J.B., Raghavan, V., Raman, R., Rana, N.P., Samothrakis, S., Spencer, J., Tamilmani, K., Tubadji, A., Walton, P. and Williams, M.D. (2021) “Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy,” International Journal of Information Management. Elsevier BV. https://doi.org/10.1016/j.ijinfomgt.2019.08.002.

El Zini, J. and Awad, M., (2023). “On the Explainability of Natural Language Processing Deep Models”. ACM Computing Surveys, 55(5), Article 103. https://doi.org/10.1145/3529755

European Commission, (2023). “Artificial Intelligence – Questions and Answers”. Available at: https://ec.europa.eu/commission/presscorner/detail/en/QANDA_21_1683 (Accessed: 21 March 2024)

Es, S., James, J., Espinosa-Anke, L., Schockaert, S., 2023. “RAGAS: Automated Evaluation of Retrieval Augmented Generation”. arXiv:2309.15217v1 [cs.CL], 26 September. https://doi.org/10.48550/arXiv.2309.15217

European Parliament. Directorate General for Parliamentary Research Services., (2020). “European framework on ethical aspects of artificial intelligence, robotics and related technologies: European added value assessment”. Publications Office, LU. https://doi.org/10.2861/94107

Enholm, I.M., Papagiannidis, E., Mikalef, P. et al., (2022). “Artificial Intelligence and Business Value: a Literature Review”. Information Systems Frontiers. doi.org/10.1007/s10796-021-10186-w.

Fatemi, S. & Hu, Y. (2023). “A Comparative Analysis of Fine-Tuned LLMs and Few-Shot Learning of LLMs for Financial Sentiment Analysis.” arXiv. https://doi.org/10.48550/ARXIV.2312.08725.

Ferrara, E. (2023a) “Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models,” arXiv [Preprint]. https://doi.org/10.48550/ARXIV.2304.03738.

Ferrara, E. (2023b) “Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies,” Sci. MDPI AG. https://doi.org/10.3390/sci6010003.

Ferrara, E. (2023c) “The Butterfly Effect in Artificial Intelligence Systems: Implications for AI Bias and Fairness,” arXiv [Preprint]. https://doi.org/10.48550/ARXIV.2307.05842.

Fridgeirsson, T.V., Ingason, H.T., Jonasson, H.I. and Jonsdottir, H. (2021) “An Authoritative Study on the Near Future Effect of Artificial Intelligence on Project Management Knowledge Areas,” Sustainability. MDPI AG. https://doi.org/10.3390/su13042345.

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., Chen, L., 2023. Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. Journal of Information Technology Case and Application Research. https://doi.org/10.1080/15228053.2023.2233814

Gartner. (2018). “Gartner says nearly half of CIOs are planning to deploy artificial intelligence”. Available at: https://www.gartner.com/en/newsroom/press-releases/2018-02-13-gartner-says-nearly-half-of-cios-are-planning-to-deploy-artificial-intelligence (Accessed 28-02-2024)

Garcia, M., (2024). “What Air Canada Lost In ‘Remarkable’ Lying AI Chatbot Case”. Forbes. Available at: https://www.forbes.com/sites/marisagarcia/2024/02/19/what-air-canada-lost-in-remarkable-lying-ai-chatbot-case/?sh=42751874696f (Accessed: 21 March 2024).

Gartner. (2019). “Gartner Says 80 Percent of Today’s Project Management Tasks Will Be Eliminated by 2030 as Artificial Intelligence Takes Over”. Available at: https://www.gartner.com/en/newsroom/press-releases/2019-03-20-gartner-says-80-percent-of-today-s-project-management (Accessed 1 March 2024)

Gill, S.S. and Kaur, R. (2023) “ChatGPT: Vision and challenges,” Internet of Things and Cyber-Physical Systems. Elsevier BV. https://doi.org/10.1016/j.iotcps.2023.05.004.

Google AI. (n.d.). Google AI PaLM 2. Available at: ai.google/discover/palm2/ (Accessed 4 March 2024)

Google Research Team. (2023) “Gemini: A Family of Highly Capable Multimodal Models.” arXiv. https://doi.org/10.48550/ARXIV.2312.11805.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) “Generative Adversarial Networks.” arXiv. https://doi.org/10.48550/ARXIV.1406.2661.

Haefner, N., Parida, V., Gassmann, O. and Wincent, J. (2023) “Implementing and scaling artificial intelligence: A review, framework, and research agenda,” Technological Forecasting and Social Change. Elsevier BV. https://doi.org/10.1016/j.techfore.2023.122878.

Hays, K & Rafieyan, D. (2024). “OpenAI is expected to release a ‘materially better’ GPT-5 for its chatbot mid-year sources say”. Business Insider. Available at: https://www.businessinsider.com/openai-launch-better-gpt-5-chatbot-2024-3 (Accessed 26 April 2024)

Hitaj, Briland, Ateniese, Giuseppe and Perez-Cruz, Fernando (2017). Deep Models Under the GAN. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security.

Hiter, S. (2023) ‘What Are AI Hallucinations and How Do They Work?’, eWeek, p. N.PAG. Available at: https://search-ebscohost-com.ludwig.lub.lu.se/login.aspx?direct=true&AuthType=ip,uid&db=bth&AN=174425223&site=eds-live&scope=site (Accessed: 22 January 2024).

Hofmann, P., Jöhnk, J., Protschky, D. and Urbach, N. (2020) “Developing Purposeful AI Use Cases – A Structured Method and Its Application in Project Management,” WI2020 Zentrale Tracks. GITO Verlag. https://doi.org/10.30844/wi_2020_a3-hofmann.

Huang, K., Zhang, F., Li, Y., Wright, S., Kidambi, V. and Manral, V. (2023) “Security and Privacy Concerns in ChatGPT,” Beyond AI. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-45282-6_11.

Hugging Face. (n.d.). “T5. In Hugging Face Transformers”. Available at: https://huggingface.co/docs/transformers/en/model_doc/t5 (Accessed 4 March 2024)

Hugging Face. (2024). “Hugging Face Hub: Models”. Available at: https://huggingface.co/ (Accessed 4 March 2024)

IBM. (2022). “AI vs. Machine Learning vs. Deep Learning vs. Neural Networks”. Available at: https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/. (Accessed 19 January 2024)

IBM. (n.d.) “Neural Networks”. Available at: https://www.ibm.com/topics/neural-networks (Accessed: 24 January 2024).

IBM. (2024). “IBM Watsonx: An AI and data platform built for business”. Available at: https://www.ibm.com/watsonx (Accessed 4 March 2024)

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A. and Fung, P. (2023) “Survey of Hallucination in Natural Language Generation,” ACM Computing Surveys. Association for Computing Machinery (ACM). Available at: https://doi.org/10.1145/3571730.

Kiefer, C. (2016) “Assessing the Quality of Unstructured Data: An Initial Overview”, LWDA. Available at: https://ceur-ws.org/Vol-1670/paper-25.pdf (Accessed: 27-02-2024)

Koller, T., Goedhart, M. & Wessels, D., 2020. Valuation: Measuring and Managing the Value of Companies. 7th ed. Hoboken, NJ: John Wiley & Sons

Kaspersen, L. (2023) “The Carbon footprint of GPT-4”, Towards Data Science. Available at: https://towardsdatascience.com/the-carbon-footprint-of-gpt-4-d6c676eb21ae (Accessed 25 March 2024)

LangChain. (2024). “Tools”. Available at: https://python.langchain.com/docs/modules/tools/ (Accessed 16 April 2024)

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. and Kiela, D. (2020) “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” arXiv. https://doi.org/10.48550/ARXIV.2005.11401.

Lin, X., Wang, W., Li, Y., Yang, S., Feng, F., Wei, Y. and Chua, T.-S. (2024) “Data-efficient Fine-tuning for LLM-based Recommendation.” arXiv. https://doi.org/10.48550/ARXIV.2401.17197.

Liu Tran, P. (2024) 'Overcoming data debt with the Data Trust Workflow', 10 January. Available at: https://validio.io/blog/overcoming-data-debt-with-the-data-trust-workflow (Accessed 19 March 2024)

Marsman, P. (2024). “pdfminer.six”. GitHub repository. Available at: https://github.com/pdfminer/pdfminer.six (Accessed 16 April 2024)

McCarthy, J., Minsky, M. L., Rochester, N. and Shannon, C. E. (2006) “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955”, AI Magazine, 27(4), p. 12. https://doi.org/10.1609/aimag.v27i4.1904

Merhi, M.I. (2023) “An evaluation of the critical success factors impacting artificial intelligence implementation,” International Journal of Information Management. Elsevier BV. https://doi.org/10.1016/j.ijinfomgt.2022.102545.

Microsoft. (2023). “LLM Fine-Tuning Recommendations”. Available at: learn.microsoft.com (Accessed 4 March 2024)

Microsoft. (2024). “Customize a model with fine-tuning”. 22 February 2024. Available at: https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/fine-tuning?tabs=turbo%2Cpython-new&pivots=programming-language-studio (Accessed 16 April 2024)

Morris, M.R., Sohl-dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C. and Legg, S. (2023) “Levels of AGI: Operationalizing Progress on the Path to AGI.” arXiv. https://doi.org/10.48550/ARXIV.2311.02462.

Mosier, K.L. and Skitka, L.J. (1999) “Automation Use and Automation Bias,” Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications. https://doi.org/10.1177/154193129904300346

Meta AI. (2024) “Introducing Meta Llama3: The most capable openly available LLM to date”. Available at: https://ai.meta.com/blog/meta-llama-3/ (Accessed 3 May 2024)

Mariani, M.M., Machado, I. and Nambisan, S. (2023) “Types of innovation and artificial intelligence: A systematic quantitative literature review and research agenda,” Journal of Business Research. Elsevier BV. Available at: https://doi.org/10.1016/j.jbusres.2022.113364

Myers, W., (1986). “Introduction to Expert Systems”. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5006506 (Accessed 19 January 2024).

Nagle, T., Redman, T.C. and Sammon, D. (2017) 'Only 3% of Companies’ Data Meets Basic Quality Standards', Harvard Business Review, 11 September. Available at: https://hbr.org/2017/09/only-3-of-companies-data-meets-basic-quality-standards (Accessed 19 March 2024).

Naughton, J. (2023). “Why AI is a disaster for the climate”. The Guardian. 23 December 2023. Available at: https://www.theguardian.com/commentisfree/2023/dec/23/ai-chat-gpt-environmental-impact-energy-carbon-intensive-technology. (Accessed 22 April 2024).

Nussey, B. (2019) 'What can you do with a megawatt-hour?', Freeing Energy. Available at: https://www.freeingenergy.com/what-is-a-megawatt-hour-of-electricity-and-what-can-you-do-with-it/ (Accessed 25 March 2024)

NVIDIA. (2023). “Generative AI Workflows Technical Brief. Nvidia”. Available at: https://docs.nvidia.com/ai-enterprise/workflows-generative-ai/0.1.0/introduction.html (Accessed 4 March 2024)

Ng, A., (2024). Agentic Design Patterns Part 1: Four AI agent strategies that improve GPT-4 and GPT-3.5 performance. Letters: Technical Insights, [online] Published 20 March 2024. Available at: https://www.deeplearning.ai/the-batch/how-agents-can-improve-llm-performance/?utm_campaign=The%20Batch&utm_medium=email&_hsenc=p2ANqtz-9XZNzY25fDulH1CsSst0m-IrmmNf2Cfde2L0n2iKKh4DgvoUKSc44PaYvF-CHbjdPg1W8uhcY95YxqIMPuXxHC-THHOw&_hsmi=302131862&utm_content=302130241&utm_source=hs_email (Accessed 18 Apr 2024)

Ng, A., (2024b). Agentic Design Patterns Part 3, Tool Use. Letters: Technical Insights [online] Published 3 Apr 2024. Available at: https://www.deeplearning.ai/the-batch/agentic-design-patterns-part-3-tool-use/ (Accessed 26 Apr 2024)

Nilsson, M. et al., (2024). Artificial Intelligence and Project Management: A Global Chapter-Led Survey. Project Management Institute. Available at: https://www.pmi.org/-/media/pmi/documents/public/pdf/artificial-intelligence/community-led-ai-and-project-management-report.pdf?rev=bca2428c1bbf4f6792f521a95333b4df (Accessed: 1 March 2024)

Nilsson, M. et al., (2024b). “Navigating AI in Project Management”. Available at: https://www.pmi-se.org/Filer/PMI/AI%20in%20PM/Case%20Study%20Report_April%202024/Navigating%20AI%20Report%2020240408.pdf?TS=638481960028941040 (Accessed: 25 April 2024)

Oguz, A. (2022) “Project Management”. MSL Academic Endeavors. Available at: https://search-ebscohost-com.ludwig.lub.lu.se/login.aspx?direct=true&AuthType=ip,uid&db=cat07147a&AN=lub.7276308&site=eds-live&scope=site (Accessed: 5 February 2024).

OpenAI. (2023). “DALL-E 2”. Available at: https://openai.com/dall-e-2 (Accessed 4 March 2024)

OpenAI. (2023a). “GPT-4”. Available at: https://openai.com/gpt-4 (Accessed 4 March 2024)

OpenAI. (2023b). “Six Strategies for Getting Better Results from Your Prompts”. Available at: platform.openai.com (Accessed 4 March 2024)

OpenAI. (2023c). “GPT-4 Technical Report.” arXiv. https://doi.org/10.48550/ARXIV.2303.08774.

OpenAI. (2023d). “Pricing”. Available at: https://openai.com/pricing (Accessed 26 April 2024)

OpenAI. (2024a). “Chat Completions response format”. Available at: https://platform.openai.com/docs/guides/text-generation/chat-completions-api (Accessed 16 April 2024)

OpenAI. (2024b). “Enterprise privacy at OpenAI”. Available at: https://openai.com/enterprise-privacy (Accessed 17 April 2024)

OpenAI. (2024c) “Hello GPT-4o”. Available at: https://openai.com/index/hello-gpt-4o/ (Accessed 17 May 2024)

Park, J.S., Barber, R., Kirlik, A. and Karahalios, K. (2019) “A Slow Algorithm Improves Users’ Assessments of the Algorithm’s Accuracy,” Proceedings of the ACM on Human-Computer Interaction. Association for Computing Machinery (ACM). https://doi.org/10.1145/3359204.

Passi, S. and Jackson, S.J. (2018) “Trust in Data Science,” Proceedings of the ACM on Human-Computer Interaction. Association for Computing Machinery (ACM). Available at: https://doi.org/10.1145/3274405.

Patel, V (2023). “Private Conversations With Google Bard Appear On Google Search, How To Stop It?”. IBT. September 29. https://www.ibtimes.co.uk/private-conversations-google-bard-appear-google-search-how-stop-it-1720059. (Accessed: 22 January 2024).

Pichai, S. (2024, February). Introducing Gemini 1.5, Google's Next-Generation AI Model [Blog post]. Google AI Blog. Available at: blog.google/technology/ai/google-gemini-next-generation-model-february-2024/ (Accessed 26 April 2024)

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M. and Dean, J. (2021) “Carbon Emissions and Large Neural Network Training.” arXiv. https://doi.org/10.48550/ARXIV.2104.10350.

People + AI Guidebook. (n.d). “Errors + Graceful Failure”. Available at: https://pair.withgoogle.com/chapter/errors-failing/. (Accessed 26 April 2024)

Petkovic, D. (2023) ‘It is Not “Accuracy vs. Explainability”—We Need Both for Trustworthy AI Systems’, IEEE Transactions on Technology and Society, Technology and Society, IEEE Transactions on, IEEE Trans. Technol. Soc, 4(1), pp. 46–53. doi:10.1109/TTS.2023.3239921

Pipino, L.L., Lee, Y.W. and Wang, R.Y. (2002) “Data quality assessment,” Communications of the ACM. Association for Computing Machinery (ACM). https://doi.org/10.1145/505248.506010.

Potasznik, A. (2023) ‘ABCs: Differentiating Algorithmic Bias, Automation Bias, and Automation Complacency’, 2023 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS), Ethics in Engineering, Science, and Technology (ETHICS), 2023 IEEE International Symposium on, pp. 1–5. doi:10.1109/ETHICS57328.2023.10155094.

PMI. (2021). “A guide to the project management body of knowledge (PMBOK® guide)”. (7th ed). Project Management Institute.

PMI. (2023) Shaping the Future of Project Management with AI. Available at: https://www.pmi.org/learning/thought-leadership/ai-impact/shaping-the-future-of-project-management-with-ai (Accessed: 24 January 2024).

Prieto, S.A., Mengiste, E.T. and García de Soto, B. (2023) “Investigating the Use of ChatGPT for the Scheduling of Construction Projects,” Buildings. MDPI AG. https://doi.org/10.3390/buildings13040857.

Ramachandran, K., Quarta, L., Schuuring, M., Kirschniak, C., Rehberg, B., & Gourévitch, A. (2024) 'The Solution to Data Management’s GenAI Problem? More GenAI.', Boston Consulting Group, 6 February. Available at: https://www.bcg.com/publications/2024/the-solution-to-data-managements-genai-problem (Accessed 2 March 2024).

Reinertsen, D.G. (2009). “The Principles of Product Development Flow: Second Generation Lean Product Development”. Celeritas Publishing.

Russell, S., & Norvig, P. (2010). “Artificial Intelligence: A Modern Approach” (3rd ed.). Pearson.

Schmidt, D. A. (2023). “Towards a Catalog of Prompt Patterns to Enhance the Discipline of Prompt Engineering” . Available at: https://www.dre.vanderbilt.edu/~schmidt/PDF/ADA_Europe_Position_Paper.pdf (Accessed 4 March 2024)

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A. and Hall, P. (2022) “Towards a standard for identifying and managing bias in artificial intelligence”. National Institute of Standards and Technology (U.S.). https://doi.org/10.6028/nist.sp.1270.

Shin, J., Tang, C., Mohati, T., Nayebi, M., Wang, S. and Hemmati, H. (2023) “Prompt Engineering or Fine Tuning: An Empirical Assessment of Large Language Models in Automated Software Engineering Tasks.” arXiv. https://doi.org/10.48550/ARXIV.2310.10508.

Shrivastav, Monika. (2023). "Barriers Related to AI Implementation in Supply Chain Management." JGIM vol.30. doi.org/10.4018/JGIM.296725

Statista (n.d) “Energy use when training large language model (LLM) models as of 2021”, Statista. Available at: https://www.statista.com/statistics/1384401/energy-use-when-training-llm-models/ (Accessed 25 March 2024)

Satell, G. (2017) 'The 4 Types of Innovation and the Problems They Solve', Harvard Business Review. Available at: https://hbr.org/2017/06/the-4-types-of-innovation-and-the-problems-they-solve (Accessed 16 April 2024).

Smith, V., Shamsabadi, A.S., Ashurst, C. and Weller, A. (2023) “Identifying and Mitigating Privacy Risks Stemming from Language Models: A Survey.” arXiv. https://doi.org/10.48550/ARXIV.2310.01424.

Solaiman, I. (2023) “The Gradient of Generative AI Release: Methods and Considerations.” arXiv. https://doi.org/10.48550/ARXIV.2302.04844 .

Sriram, A. (2023). ChatGPT-owner OpenAI fixes 'significant issue' exposing user chat titles. Reuters. March 22. Available at: https://www.reuters.com/technology/chatgpt-owner-openai-fixes-significant-issue-exposing-user-chat-titles-2023-03-22/. (Accessed: 22 January 2024).

Strobelt, H., Webson, A., Sanh, V., Hoover, B., Beyer, J., Pfister, H. and Rush, A.M., (2023). “Interactive and Visual Prompt Engineering for Ad-hoc Task Adaptation with Large Language Models”. IEEE Transactions on Visualization and Computer Graphics, 29(1). Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9908590 (Accessed 23 January 2024)

Strubell, E., Ganesh, A. and McCallum, A. (2019) “Energy and Policy Considerations for Deep Learning in NLP.” arXiv. https://doi.org/10.48550/ARXIV.1906.02243.

Studer, S., Bui, T.B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S. and Müller, K.-R. (2021) “Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology,” Machine Learning and Knowledge Extraction. MDPI AG. https://doi.org/10.3390/make3020020.

Takagi, Nilton & Varajão, João. (2020). Success Management and the Project Management Body of Knowledge (PMBOK): An Integrated Perspective -research-in-progress.

Taleb, I., Serhani, M.A. and Dssouli, R. (2018) 'Big Data Quality Assessment Model for Unstructured Data', 2018 International Conference on Innovations in Information Technology (IIT), Al Ain, United Arab Emirates, pp. 69-74. doi: 10.1109/INNOVATIONS.2018.8605945.

Taboada, I., Daneshpajouh, A., Toledo, N. and de Vass, T. (2023) “Artificial Intelligence Enabled Project Management: A Systematic Literature Review,” Applied Sciences. MDPI AG. Available at: https://doi.org/10.3390/app13085014.

TRG Datacenters (2023) 'AI Chatbots: Energy usage of 2023's most popular chatbots (so far)', TRG Datacenters. Available at:

https://www.trgdatacenters.com/resource/ai-chatbots-energy-usage-of-2023s-most-popular-chatbots-so-far/ (Accessed 25 March 2024).

Tobin, Josh. 2024. "Evaluating LLM-based Applications" YouTube. Available at: https://www.youtube.com/watch?v=2CIIQ5KZWUM (Accessed: 18 March 2024).

Tomlinson, B., Black, R.W., Patterson, D.J. and Torrance, A.W. (2024) "The carbon emissions of writing and illustrating are lower for AI than for humans," Scientific Reports. Springer Science and Business Media LLC. https://doi.org/10.1038/s41598-024-54271-x.

Vakilzadeh, S.A., PourAhmad Ghalejoogh, S. and Hatami, M. (2023) 'Evaluating the Potential of Large Language Model AI as Project Management Assistants: A Comparative Simulation to Evaluate GPT-3.5, GPT-4, and Google-Bard Ability to pass the PMI's PMP test'.. Available at: https://ssrn.com/abstract=4568800 (Accessed: 24 January 2024).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., (2017). Attention Is All You Need. In: Advances in Neural Information Processing Systems. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (Accessed 19 January 2024).

Wang, Q. (2019) 'How to apply AI technology in Project Management', PM World Journal, VIII(III), April. Available at: www.pmworldjournal.com (Accessed: 24 January 2024).

Weisz, J.D., He, J., Muller, M., Hoefer, G., Miles, R. and Geyer, W. (2024) "Design Principles for Generative AI Applications," arXiv [Preprint]. https://doi.org/10.48550/ARXIV.2401.14484.

Weisz, J.D., Muller, M., He, J. and Houde, S. (2023) "Toward General Design Principles for Generative AI Applications." arXiv. https://doi.org/10.48550/ARXIV.2301.05578.

Weng, J. (2023) Implementing Generative AI Tools in Project Management. [White paper] New York University. Available at: https://archive.nyu.edu/handle/2451/69531 (Accessed: 24 January 2024).

Wiechork, K. and Charão, A. (2021) "Automated Data Extraction from PDF Documents: Application to Large Sets of Educational Tests," Proceedings of the 23rd International Conference on Enterprise Information Systems. 23rd International

Conference on Enterprise Information Systems, SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0010524503590366.

Wikipedia contributors, (2024). Computer performance by orders of magnitude. Available at: https://en.wikipedia.org/wiki/Computer_performance_by_orders_of_magnitude (Accessed 23 January 2024).

Weber, M., Limmer, N. & Weking, J., 2022. Where to Start with AI?—Identifying and Prioritizing Use Cases for Health Insurance. Proceedings of the 55th Hawaii International Conference on System Sciences, [online] Available at: https://hdl.handle.net/10125/79818 (Accessed 24 April 2024)

Wirth, R. and Hipp, J., (2000). "CRISP-DM: Towards a Standard Process Model for Data Mining". DaimlerChrysler Research & Technology FT3/KL; Wilhelm-Schickard-Institute, University of Tübingen. Available at: https://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf (Accessed 22 January 2024).

Zarifhonarvar, A. (2023) "Economics of ChatGPT: a labor market view on the occupational impact of artificial intelligence," Journal of Electronic Business & Digital Economics. Emerald. https://doi.org/10.1108/jebde-10-2023-0021.

Zaman, G., Mahdin, H., Hussain, K. and Atta-ur-Rahman (2020) "Information Extraction from Semi and Unstructured Data Sources: A Systematic Literature Review," https://doi.org/10.24507/icicel.14.06.593.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D. and Du, M., (2024). "Explainability for Large Language Models: A Survey". ACM Transactions on Intelligent Systems and Technology (TIST), 15(2), Article 20. https://doi.org/10.1145/3639372

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E. and Stoica, I. (2023) "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena." arXiv. https://doi.org/10.48550/ARXIV.2306.05685.

Ångström, R. C., Björn, M., Dahlander, L., Mähring, M., & Wallin, M. W. (2023). "Getting AI Implementation Right: Insights from a Global Survey". California Management Review. doi.org/10.1177/00081256231190430

Appendix A - Sample training data for fine tuning

TEXT:
QUOTATION
2024-03-05
Date
1
Page

Quotation No

58733

Your reference

Terms of delivery

CIF

Delivery time

Period of validity

90 Days

Terms of payment

Net 90 days

Our reference

Customer

GreenTech Solutions Ltd.

789 Eco Park

Green Valley, 83000

Thank you for your request. We are pleased to give you the following quotation.

Item No

520489

Solar Panel 300W

Specs: SP300-2024 Model, Monocrystalline, 20% Efficiency

Eng 304957-63-21

520490

Wind Turbine 5kW

Specs: WT5K-2024 Model, 3 Blades, 6m Rotor Diameter

Eng 304958-63-22

520491

Battery Storage 10kWh

Specs: BS10-2024 Model, Lithium-Ion, 10 Years Warranty

Eng 304959-63-23

Mtrl authorization

Lead working days

Qty per year

Qty per batch

Unit price

0,00

0,00

0,00

0,00

0,00

0,00

200

5,00

5,00

245.00 USD

150

2,00

2,00

8,750.00 USD

100

4,00

4,00

5,200.00 USD

Prices are subject to change based on global market trends and currency exchange rates.

Metal and currency adjustments
Above prices are based on the current exchange rate as of 2024-03-05
1,00 USD = 0.92 EUR

Currency adjustment according to actual exchange rate at order date or calculation date in case of modification.

Please send your order to: sales@greentechsolutions.com

Other conditions

We look forward to receiving your order.
If you have any questions, please don't hesitate
to contact me.

Best regards,

.@greentechsolutions.com

Issued by GreenTech Solutions Ltd.

www.greentechsolutions.com

ASSIGNMENTS=

Think through the TEXT carefully and list systematically in a csv format the Quotation number, Supplier, Item ID, Description, Quotation date, Period of validity, Price, Currency, Quantity per batch, Exchange rate

RESPONSE:
58733;GreenTech Solutions;520489;Solar Panel 300W;2024-03-05;90;245.00;USD;5;0.92;;;
58733;GreenTech Solutions;520490;Wind Turbine 5kW;2024-03-05;90;8750.00;USD;2;0.92;;;

58733;GreenTech Solutions;520491;Battery Storage 10kWh;2024-03-05;90;5200.00;USD;4;0.92;;;

Appendix B - Cost calculations for financial model of budget demo

Cost savings from current process

Hours saved on task/ project	25
Projects/year/project leader	6
Number of project leaders	10
Total hours on task/year	1500
Cost/ project leader/ hour (€)	120
Total cost/ year (€)	**180000**

Man hour costs of development

Build	
Time to build (months)	3
Hours/month	160
Time to build (hours)	480
Cost/hour (€)	120
Total cost to build	**57600**
Maintain	
Hours to maintain/week	2
Weeks/year	48
Cost/hour	120
Total maintenance cost	**11520**

Development costs of running the model

	Fine tuned GPT-3.5 Turbo	GPT-4
Training		

Tokens/PDF	15000	-
Number of PDF for training	1000	-
Total tokens for training	**15000000**	-
Cost/token for training (/1M tokens) ($)	8	-
Total training cost ($)	**120,00**	-
Additional cost of testing etc ($)	120,00	120
Total development costs ($)	240,00	120,00
Total development costs (€)	**220,8**	**110,4**

Runtime costs of using the model

	Fine tuned GPT-3.5 turbo	GPT-4
Input		
Tokens/PDF	15000	15000
Number of PDFs/input (avg)	100	100
Number of input/project	2	2
Projects/year	80	80
Total input tokens/year	**240 000 000**	**240 000 000**
Cost/input token (/1M tokens) ($)	3	30
Total input cost per year	**720**	**7200**
Output		
Number of tokens/output	30	30
Number of outputs/call	100	100
Number of calls/year	160	160
Total output tokens/year	**480000**	**480000**
Cost/output token	6	60
Total output cost per year	**2,88**	**28,8**
Total runtime costs/year ($)	722,88	7228,8
Total runtime costs/year (€)	**665,0496**	**6650,496**