

Seeing the Unseen

AI-Driven Discrepancy Detection to Foster Improved
Decision-Making and Supply Chain Resilience

A Case Study of PipeChain SCM

MTTM05 Degree Project

Authors:

Emma Iveberg & Erik Ekstrand

Supervisor at LTH: Sandeep Jagtap

Supervisor at PipeChain SCM: Charlotte Widstrand

Examiner: Joakim Kembro

Division of Engineering Logistics

Faculty of Engineering - LTH

Lund University

Lund, June 2026

Acknowledgments

This master thesis was conducted in collaboration with the Division of Engineering Logistics at LTH, Lund University. The completion of this thesis marks the completion of five years of academic studies for both authors, with Emma graduating in Industrial Engineering and Management and Erik in Mechanical Engineering.

First and foremost, we would like to express our sincere gratitude to our academic supervisor, Sandeep Jagtap, for his invaluable guidance, support, and constructive feedback throughout the entire thesis process. His insights and encouragement have been instrumental to the developments and completion of this work, and this thesis would not have been possible without his support.

We would also like to extend our deepest appreciation to PipeChain SCM for welcoming us and providing us with the opportunity to collaborate with the company. Throughout the thesis project, the company consistently supported us with data, tools, and valuable expertise, enabling the work to progress efficiently. We would especially like to thank Jens Dremo for making this collaboration possible, as well as Lennart Hagberg and Daniel Larsson for their support in the practical aspects of the project.

Furthermore, we would like to express our genuine appreciation to our supervisor at PipeChain SCM, Charlotte Widstrand. From the very beginning, Charlotte has been an invaluable source of support in both practical and academic matters. Whether assisting us in navigating daily routines as the office, arranging interviews, or inviting us to an insightful conference, her guidance and engagement have greatly contributed to the progress of this thesis. Our weekly meetings have been particularly valuable in ensuring steady progress throughout the project.

Lastly, we would like to extend a heartfelt thank you to our families and loved ones for their continuous encouragements and support throughout these five years of study.

Abstract

Background

Supply Chain Management is becoming increasingly more important to organizations in the post Covid-19 era. Managing Supply Chains is also moving towards finding data-oriented ways of working, and companies are making substantial investments in getting digital information systems to integrate with their ways of working. Simultaneously, Artificial Intelligence has become a well-discussed topic over the past few years, with the introduction of large language models and chatbots to the public. There is a massive interest across many industries to find ways in which generative AI can be used to create value, and supply chain companies are no different.

Problem Definition and Purpose

When managing their supply chains, actors often have access to a number of different systems to monitor operations, e.g. ERP-, customer relationship management- or supplier relationship management-systems. These systems displays data to support system users in their decision making. However, it takes time and effort to reliably be able to learn and navigate from different systems. According to research, one of AI's main strengths is its ability to perform analysis on big data sets. This thesis aims to explore if generative AI can be used to interpret, analyze and present its found insights in supply chain processes to a system user, and thereby making it easier to gather insights from available data.

Method

The thesis deploys a case study, using an abductive research approach. Initially a literature review and a mapping of a case-specific process to establish a foundation of what generative AI will have to interpret and present. Thereafter, current frameworks for AI-prompting are looked into, and then turned into a new one fitting for the case. Finally, the framework is applied to the case, and generated output is evaluated to identify strengths and weaknesses of generative AI's analysis.

Conclusions

The study finds that generative AI can be able to interpret, structure and visualize data in a reliable way in a supply chain setting. The quality and procedure of the analysis is highly dependent on which generative AI model is used. Between the evaluated models, Claude was the best performing, closely followed by ChatGPT, while Gemini and Copilot showed weaker results. Furthermore, all models achieved worse overall output in a test environment where two metrics was to be analyzed compared to only looking at one.

Key Words: Supply Chain Management, Artificial Intelligence, Prompt Engineering, Large Language Model Evaluation, Deviation Detection

Abbreviations

AI	Artificial Intelligence
CA	Calculation Accuracy
CDR	Consistency of Deviation Reliability
CoT	Chain-of-Thought
D	Demanded
DP	Delivery Precision
DPA	Deviation Precision Accuracy
DSR	Design Science Research
E1	Environment 1
E2	Environment 2
FA	Forecast Accuracy
FB	Forecast Bias
FC	First Confirmed
GenAI	Generative AI
KPI	Key Performance Indicator
LC	Last Confirmed
LLM	Large Language Model
ML	Machine Learning
NLP	Natural Language Processing
PC	PipeChain
PC SCM	PipeChain SCM
PMS	Performance Measurement System
S	Sensitivity
SCM	Supply Chain Management
SCOR	Supply Chain Operations Reference
SL	Service Level
SRM	Supplier Relationship Management
T1	Test 1
T2	Test 2
T3	Test 3
T4	Test 4
T5	Test 5
T6	Test 6
VMI	Vendor Managed Inventory

List of Figures

1.1	<i>A breakdown of the offerings of the system PC, including what each part offers and tracks.</i>	3
2.1	<i>The balanced approach model constructed by Golicic et al. (2005), illustrating the method of combining the inductive and deductive approaches.</i>	7
2.2	<i>DSR framework according to Brocke et al. (2020)</i>	9
2.3	<i>The single-embedded case study design adopted in this study, developed by Yin (2018).</i>	11
2.4	<i>An overview of the abductive research approach of the thesis</i>	17
3.1	<i>Visual representation of the theoretical framework and its key components</i>	18
3.2	<i>A conceptual supply chain according to Olhager (2019)</i>	19
4.1	<i>Information flows from replenishment process in the system.</i>	40
4.2	<i>Frequently mentioned metrics in conducted interviews.</i>	43
4.3	<i>Frequently mentioned areas of value creation in conducted interviews.</i>	44
4.4	<i>Frequently mentioned usage areas in conducted interviews.</i>	45
6.1	<i>An adaptation of the SAFE-PE methodology presented by (Laeq and Saleem, 2025), adapted to this thesis</i>	59
6.2	<i>Overview of evaluation of test-results based on the SAFE-PE framework.</i>	61
6.3	<i>The ten suppliers with the most deviating rows in E1.</i>	66
6.4	<i>The ten products with the most deviating rows in E2.</i>	67
6.5	<i>The number of presented suppliers by Gemini, ChatGPT and Claude, for each of the five iterations in T4, E1 along with the expected number.</i>	70
6.6	<i>All presented suppliers and their frequency of being deemed one of the five most problematic suppliers by Gemini, ChatGPT and Claude in T4, E1.</i>	70
6.7	<i>The number of presented products by Gemini, ChatGPT and Claude, for each of the five iterations in T4, E2 along with the expected number.</i>	71
6.8	<i>The presented products by Gemini, ChatGPT and Claude during T4, E2 and their frequency of being deemed one of the five most problematic issues.</i>	71
6.9	<i>The number of presented suppliers each iteration by Gemini, ChatGPT and Claude for T5, E1, along with the expected number.</i>	72
6.10	<i>The frequency of presented suppliers by Gemini, ChatGPT and Claude for T5, E1.</i>	73
6.11	<i>The number of presented products by Gemini, ChatGPT and Claude for each iteration with a modified prompt in T5, E2, along with the expected number.</i>	73
6.12	<i>The presented products by Gemini, ChatGPT and Claude during T5, E2 and their frequency of being deemed one of the five most problematic issues.</i>	74

6.13	<i>Cross-model comparison displaying all test results from applying the SAFE-PE framework to the adjusted prompt-framework in E1</i>	75
6.14	<i>Cross-model comparison displaying all test results from applying the SAFE-PE framework to the adjusted prompt-framework in E2</i>	75

List of Tables

3.1	Comparison of prompting techniques	31
5.1	Search words used for exploring existing frameworks for prompting genAI on Finn.	48
5.2	Search words used for exploring existing frameworks for prompting genAI on Scopus.	48
5.3	Overview of prompt-based frameworks	49
5.3	Overview of prompt-based frameworks	50
5.4	The theoretical prompting framework with proposed key components of the prompt	53
5.5	The adjusted prompting framework with proposed key components of the prompt	57
6.1	Conducted tests using the SAFE-PE framework and the research questions connected to each test.	61
6.2	Expert assessment areas for qualitative prompt evaluation.	63
6.3	Versions of the genAI models used for testing and their originators.	65
6.4	Summary of the number of generated values and results from T1 and T2 in both test environments for all genAI models.	67
6.5	The average results from the assessments conducted in T3 for E2.	68
6.6	The average results from the assessments conducted in T3 for E2.	69
6.7	Summary of the results from T4 and T5 in both test environments for all genAI models.	69
A.1	Interviewees, their role and system experience	100
C.1	Screening of existing prompting frameworks in relation to PipeChain SCM requirements	103
F.1	The results from the assessments conducted in T3 on the output generated from ChatGPT in E1.	117
F.2	The results from the assessments conducted in T3 on the output generated from ChatGPT in E2.	117
F.3	The results from the assessments conducted in T3 on the output generated from Claude in E1.	117
F.4	The results from the assessments conducted in T3 on the output generated from Claude in E2.	117
F.5	The results from the assessments conducted in T3 on the output generated from Gemini in E1.	118
F.6	The results from the assessments conducted in T3 on the output generated from Gemini in E2.	118
G.1	Prompt framework for T1 and T4 in E1	119
G.1	Prompt framework for T1 and T4 in E1	120
G.2	Prompt framework for T1, T3, and T4 in E2	121
G.2	Prompt framework for T1, T3, and T4 in E2	122

G.3	Prompt framework for T2 in E1	123
G.3	Prompt framework for T2 in E1	124
G.4	Prompt framework for T2 in E2	125
G.4	Prompt framework for T2 in E2	126
G.5	Adjusted prompt for E1, first iteration of T5	127
G.6	Adjusted prompt for E1, second iteration of T5	128
G.7	Adjusted prompt for E1, third iteration of T5	129
G.8	Adjusted prompt for E2, first iteration of T5	130
G.9	Adjusted prompt for E2, second iteration of T5	131
G.10	Adjusted prompt for E2, third iteration of T5	132

Table of Content

1	Introduction	1
1.1	Background	1
1.2	PipeChain	2
1.3	Problem Statement	3
1.4	Purpose and Research questions	3
1.5	Scope, Boundaries, and Delimitations	4
1.6	Non-disclosure Agreement	5
1.7	Thesis Outline	5
2	Methodology	7
2.1	Research Approach	7
2.2	Research Strategy	8
2.3	Data Collection	11
2.4	Data Analysis	14
2.5	Research Quality	15
2.6	Summary of Methodology	16
3	Literature Review	18
3.1	Supply Chain Management	18
3.2	Artificial Intelligence	26
3.3	Artificial Intelligence Within the Supply Chain Management Context	34
4	Mapping of PipeChain SCM's Inbound Process	38
4.1	Purpose of Interviews	38
4.2	Interviewees	38
4.3	Case-Specific Process	39
4.4	Available data	41
4.5	Identifying Points of Interest	42
4.6	Takeaways	45
5	Constructing Foundation for Artificial Intelligence Prompting	47
5.1	Framework Development Approach	47
5.2	Existing Frameworks	47
5.3	Construction of Theoretical Framework	51
5.4	Applying the Theoretical Framework	53
5.5	Restructuring the Framework	54
5.6	Final Framework	57
6	Evaluation of Framework	59
6.1	Testing	59
6.2	Results	66
7	Discussion	77

7.1	Phase 1 - Quantitative Metrics	77
7.2	Phase 2 - Qualitative Metrics	78
7.3	Phase 3 - Robustness Metrics	79
7.4	Implications of the Evaluation	81
8	Conclusion	83
8.1	Addressing the Purpose and Research Questions	83
8.2	Research contribution	85
8.3	Limitations and Suggestions for Future Research	86
	Appendix	99
A	Interviewees	100
B	Interview Questions	101
C	Screening Table	103
D	Output from Theoretical Framework	104
E	Output from Adjusted Framework	108
F	Results from T3	117
G	Prompts	119

1 Introduction

This chapter provides a background to the thesis and the topics that are explored, stating its relevance, and providing the reader with an overall understanding. A brief presentation of the company is also provided, followed by the formulation of the problem. The purpose of the theses follows next, alongside the research questions that the thesis aims to address, after which the section is concluded by a brief overview of the remaining chapters.

1.1 Background

Digitalization in supply chain operations has created new possibilities for organizations to share information, visualize processes and track supplier performance in real time. Information systems offer a wide range of opportunities to track operations, and have become a crucial support tool for employees when making decisions throughout supply chains. Implementing artificial intelligence (AI), machine learning (ML) and large language models (LLM) to support decision making currently sparks interest across various industries.

Many companies are investigating opportunities to involve AI and ML technologies to further increase performance, with the interest in these technologies only increasing (Helo and Hao, 2021). According to Khokrale (2025), AI is seen as a strategic tool by which an organization can manage the increasing complexity that follows the accelerating digital development of supply chain operations. One of AI's main strengths is its ability to recognize patterns, especially when presented with larger data sets. (Dumitrascu et al., 2020). Therefore, integrating AI into supply chain information systems, in which huge amounts of data is collected on a daily basis, presents an opportunity in automating data management and presentation. These technologies allow for companies to anticipate and respond to disturbances in a more effective and efficient way (Riad et al., 2024).

Although AI is sometimes perceived as a relatively recent development, it has been applied within supply chain management (SCM) for several years. However, due to its ability to address complex problems in a human-like manner, it has attracted considerable attention in recent years (Ferreira et al., 2025). Furthermore, the emergence of publicly accessible LLMs has raised greater awareness of the potential of the technology, sparking interest in both industry and academia.

An increasing number of organizations are integrating AI across key parts of their value chains, as it has been shown to generate significant competitive advantages (Dash et al., 2019). Ferreira et al. (2025) highlights the strategic importance of AI in the increasingly dynamic and globalized supply chain environment, emphasizing its potential to enhance productivity through the automation and digitalization of operations. To remain competitive, organizations must adopt such technologies, as they provide the foundation for managing the growing complexity of modern supply chain systems (Genetti et al., 2025).

However, implementing digital systems and technologies into everyday operations may entail challenges in terms of value creation. Major issues include getting decision-makers to trust and embrace new tools, as well as the loss of authority. Hence, when introducing new systems it is crucial with a strong change management and a well stated purpose of how the technology is aimed at helping employees (Booyse and Scheepers, 2024). This thesis aspires to investigate whether the system's collected data can, through generative artificial intelligence (genAI), be presented along with recommended actions to users, and thereby support decision-making.

In summary, whilst long existing SCM tools, such as enterprise resource planning systems, have become progressively more automated and enabled the digitalization of many processes, there is a perceived lack of a holistic solution allowing for the many fragmented systems to rapidly, rationally, and effectively respond to changes within the supply chain. To get an understanding of these system is time consuming and demands great effort from stakeholders. Here, AI-supported analysis presents an opportunity to decrease this challenge. Although a lot of emerging research exists, extensive studies on the role of AI in SCM remain relatively scarce due to a limited understanding of this phenomenon. (Helo and Hao, 2021)

1.2 PipeChain

Founded in Sweden in the 1970s, PipeChain AB operated in the supply chain landscape during the 1990's and has since evolved into a cloud-based service company (PipeChain, n.d.). Today, the company operates as a software as a service provider, offering customers both a standardized plug-and-go in-house developed system, but also the flexibility to tailor the system to the customer's own business processes. Within the digital information environment, PipeChain AB enables automated information sharing as the system connects supply chain actors both upstream and downstream, with vendor managed inventory (VMI) and traceability being common applications. Following the acquisition of Tyringe Konsult AB in 2022, the company operates through two business areas: PipeChain Networks and PipeChain SCM (PC SCM).

Among the tools and solutions that PC SCM offers is a system called PipeChain (PC) that covers demand management, order management, delivery management, and invoice management. In addition to these areas, document and specification management can also be incorporated. Within demand management, the information flow consists of schedules and forecasts, and the system continuously tracks the forecast accuracy and bias. Order management handles purchase orders, order changes, and order responses. Here, the system tracks the requested delivery date and the suppliers confirmed delivery date. In delivery management, information sharing consists of dispatch advice and goods receipt date, where the system keeps track of the latest confirmed delivery date, as well as the goods received date. Lastly, in invoice management, invoices are shared. This is illustrated in Figure 1.1 below.

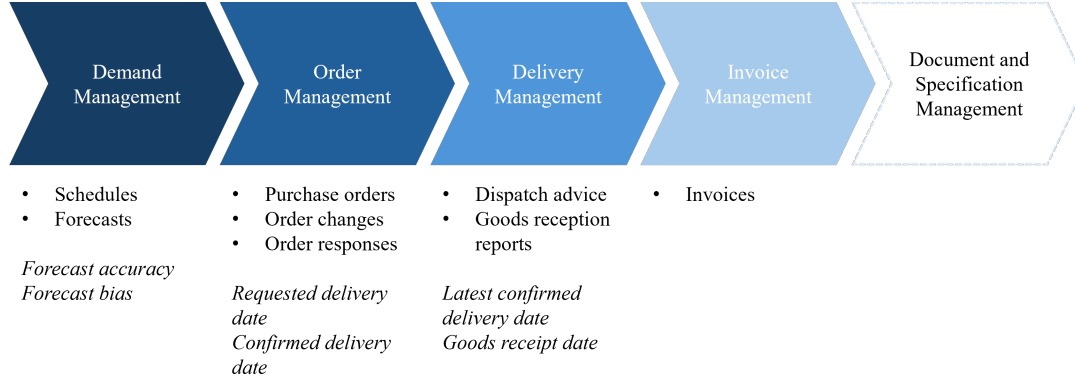


Figure 1.1: A breakdown of the offerings of the system PC, including what each part offers and tracks.

Although automation is central to the PC value proposition, the system still requires manual user interaction. Clients have highlighted a lack of data visualization capabilities, which limits their ability to effectively interpret shared information. As a result, PC SCM has expressed interest in developing new solutions that enhance data tracking and analysis within the platform, with AI integration identified as particularly interesting to improve user experience.

1.3 Problem Statement

As mentioned, PC SCM's system currently provides limited data visualization capabilities, relying primarily on standardized tables displaying KPIs and data through numerical values in predefined cells. As a result, users must perform extensive filtering and manual data handling to identify relevant key performance indicators (KPIs) and performance metrics, a process that is time-consuming and inefficient. Although the dashboard can highlight changes in the KPIs, users are required to manually examine the underlying data to detect deviations. When navigating large volumes of data to detect deviations, AI presents a promising solution that offers the potential for more efficient workflows as well as dynamic and adaptive visualization.

In the broader context of digital supply chains, effective data management, interpretation, and visualization are critical to operational decision-making. Although digital systems generate vast amounts of data, many organizations struggle to effectively provide the user with a clear overview of the available data, and this way facilitate analysis and possible issue identification. This increases the risk of performance-affecting events going undetected. Consequently, the limitations of users not possessing expertise of digital systems restrict the ability of organizations to respond proactively to changes in supply chain performance.

1.4 Purpose and Research questions

The main purpose of this thesis is to explore how genAI can be integrated into digital supply chain environments to create value for system users. Specifically, the study aims to investigate how genAI can support operational decision-making by

automating the interpretation of KPIs and this way proactively identify potential issues and improve decision support. By enabling a higher level of automation in data handling and interpretation, such solutions may improve decision-making quality, increase operational efficiency, and allow supply chain companies to extract deeper insights from KPI data. To address this, the following research questions have been formulated, which aim to break down the overarching purpose into specific measurable objectives.

RQ1: How can genAI-driven analysis be used to interpret changes in KPIs and in this way identify real time issues?

This question addresses the general question of whether genAI can be implemented in operational processes to streamline the transition from static data monitoring to proactive risk management. By testing genAI’s ability to recognize patterns in data, the study explores whether automated analysis can detect potential disruptions, enabling users to more efficiently anticipate risk before they impact operations.

RQ2: Can genAI be used to prioritize and present KPIs based on operational context and decision-making needs?

This question explores the ability of genAI finding large amount of issues in data, as well as its capacity to classify these in order of cruciality. Presenting too many issues to a system user might be overwhelming and impede the focus of continuous improvement. However, not presenting enough crucial issues risk damaging supply chain performance. Therefore, testing genAI’s ability to present the user with an adequate number of issues to focus its attention on may improve the efficiency and overall performance of the supply chain, and at the same time address PC SCM’s need for better data visualization and less manual filtering.

RQ3: To what extent can genAI find causes of shifts in KPI-values and give recommended actions?

This question moves beyond observation into root-cause analysis. The goal is to determine whether genAI can provide reliable error messages or decision logs that explain why a change occurred and recommend specific actions, thus increasing user trust in the system.

The research questions will primarily be addressed through the construction of a prompting model and then a practical testing of this using several genAI models and historical data provided by PC SCM. Additionally, a literature study will be executed to gather relevant theory, and interviews will be conducted with PC SCM system users with the objective of laying the foundation for creating a logic structure that the AI will follow. The methodology will be explained in further detail in Chapter 2.

1.5 Scope, Boundaries, and Delimitations

Existing research suggests that AI-assisted decision-making can improve supply chain performance (Larson et al., 2023). However, decision-making occurs at multiple levels within organizations and varies in scope, impact, and strategic importance.

This thesis focuses on operational and tactical decisions that influence supply chain performance on a short to mid-term perspective. It adopts a single embedded case study, centered on PC SCM and specifically on the inbound process in PC, which is further described in Chapter 4.

The study integrates key areas, including the application of AI in supply chain contexts, KPI prioritization, and digitally supported decision-making. To ensure a feasible and well-defined scope, no practical implementation of data visualization will be conducted. Instead, the output will consist of a AI-generated presentation containing explanations of KPI deviations, including identified root-causes and potential actions. Furthermore, the literature review establishes characteristics of genAI, and together with qualitative data from interviews, this sets the boundaries for the potential use case. The empirical testing is limited to the interactions within the relationships between supply chain actors.

Although the implementation of AI to enhance human labor efficiency is often discussed from an ethical perspective (Danilevskyi et al., 2025), such considerations are beyond the scope of this thesis. Instead, the focus lies on evaluating the extent to which AI can interpret and explain data, and thereby assess the potential for automating manual data handling using current technologies. Similarly, the process of implementing genAI analysis into companies' daily operations and challenges in regards to this, will not be discussed.

Moreover, the aspects of latency and computational cost is excluded from the analysis and evaluation in this thesis. Instead, emphasis is solely put on quantitative computational capacity and quality of generated output.

These delimitations were made to ensure a feasible scope within the timeframe of the thesis and to maintain focus on evaluating the quality and analytical capabilities of genAI-generated explanations, rather than broader implementation, ethical, or technical considerations.

1.6 Non-disclosure Agreement

As previously stated, this thesis relies on data provided by PC SCM. The dataset consists of real customer data, and access is therefore subject to a non-disclosure agreement. Consequently, all suppliers and products presented in this report have been anonymized. These modifications preserve the underlying ratios and relationships within the data to ensure that the results remain representative while safeguarding the confidentiality of the companies involved.

1.7 Thesis Outline

This section represents an overview of each of the following chapters of this thesis, following this introductory chapter:

Chapter 2: Methodology

The second chapter offers an overview of the chosen research methodology adopted in this thesis. Beginning with a presentation of the research approach and strategy the case study is then further discussed. After this, the data collection and analysis methods are presented before the chapter is concluded with a reasoning regarding the research quality.

Chapter 3: Literature Review

The third chapter reviews existing literature related to the two main areas of this thesis: SCM and AI. The objective is to establish the theoretical foundation for the thesis by defining key concepts within these two areas, before concluding the chapter by combining the two areas and identifying a research gap motivating the study.

Chapter 4: Mapping of PipeChain SCM's Inbound Process

The fourth chapter presents and summarizes the information and insights collected from conducted interviews with key employees at the case company as well as a customer. It introduces and describes the inbound process which is the main focus of this study.

Chapter 5: Constructing Foundation for Artificial Intelligence Prompting

The fifth chapter introduces the prompting framework used in the thesis. By first examining existing frameworks for prompting genAI, the chapter then presents first a theoretical, then an adjusted, prompting framework accommodated to data analysis in a supply chain context.

Chapter 6: Evaluation of Framework

The sixth chapter involves practical testing of the framework presented in the previous chapter. By first applying an existing prompt evaluation framework, the prompt is assessed through six different tests, evaluating quantitative, qualitative, and robustness metrics.

Chapter 7: Discussion

The seventh chapter discusses the empirical findings presented in the chapter prior to this. In line with the evaluation framework, the discussion is divided into quantitative, qualitative, and robustness analyses before concluding with an overall evaluation of the opportunities and challenges.

Chapter 8: Conclusion

The eighth chapter begins by revisiting the purpose of the thesis and then presents how the three research questions have been addressed based on the empirical findings. Thereafter, the chapter discusses limitations of the research, and states suggestions for further research.

2 Methodology

This chapter presents the working methods to be used with the aim to address the previously stated research questions. Beginning with the overall research approach, the chapter then provides an overview of the research strategy, leading on to the data collection. This section briefly discusses the structure of the literature review as well as the gathering of both quantitative and qualitative data. Then, the chapter expands on the data analysis, again divided into a quantitative and qualitative perspective. A discussion regarding the research quality and a short summary finishes of the chapter.

2.1 Research Approach

The choice of research methodology depends on the characteristics and purpose of the thesis. According to Höst et al. (2006), research can generally be classified into four main purposes: descriptive, exploratory, explanatory, and problem solving. A descriptive approach aims to describe and document how a specific phenomenon occurs or is structured. An exploratory approach is used when knowledge about a phenomenon is limited and seeks to generate insights and deeper understanding of the subject area. An explanatory approach focuses on identifying and analyzing causal relationships to explain why a phenomenon occurs. Finally, a problem solving approach aims to develop solutions to an identified problem.

Furthermore, the research approach can be either deductive or inductive (Golicic et al., 2005; Saunders et al., 2006). A deductive approach involves testing existing theories and is commonly associated with quantitative data, whereas an inductive approach aims to develop theory based on data analysis and typically emphasizes qualitative data. However, Golicic et al. (2005) argue that combining deductive and inductive reasoning is often preferable, resulting in a more balanced research approach, commonly referred to as an abductive approach (Saunders et al., 2006). The two approaches are illustrated in Figure 2.1 below.

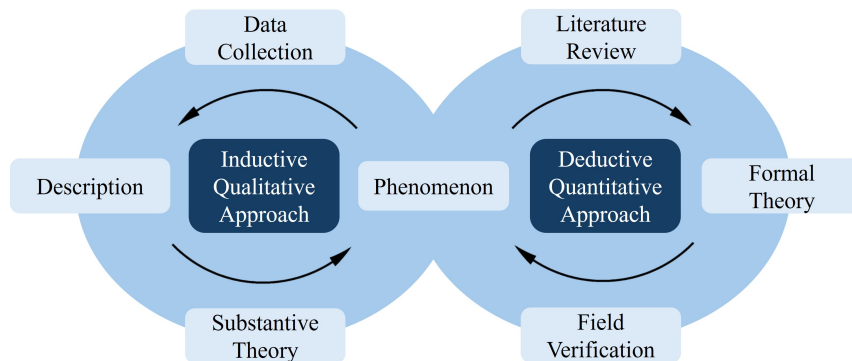


Figure 2.1: *The balanced approach model constructed by Golicic et al. (2005), illustrating the method of combining the inductive and deductive approaches.*

This thesis primarily adopts a problem solving approach. To accommodate both theory development and empirical analysis, an abductive research approach is employed, allowing for a flexible iterative movement between existing literature and case-specific data. While incorporating qualitative insights, the study places greater emphasis on quantitative analysis. This approach is particularly suitable in evaluating genAI output since relevance of qualitative output is highly dependent on quantitative output being reliable. The research is structured around three main components: first, an investigation of the specific supply chain process through interviews; second, a framework for prompting genAI is developed; and finally, the framework is applied to real case data and the generated output is evaluated.

2.2 Research Strategy

As argued by Golicic et al. (2005), no single research strategy can be considered universally superior, as each approach entails specific strengths and limitations. Quantitative methods typically offer a high degree of control and external validity, whereas qualitative methods provide greater depth and internal validity. Given the increasing complexity of supply chain environments, Golicic et al. (2005) further argue that combining multiple approaches is often preferable in order to more accurately capture and describe the phenomenon under study.

Höst et al. (2006) distinguish four primary research strategies: surveys, case studies, experiments, and action research. Surveys aim to map broad phenomena through an as-is analysis, while case studies involve an in-depth analysis of one or several cases. Experimental research focuses on comparing phenomena by isolating and manipulating influencing factors, whereas action research entails a structured and well-documented intervention aimed at solving a specific problem. According to Saunders et al. (2006), case studies are particularly suitable when the research aims to explore complex phenomena within their real-world context. Rather than isolating variables in controlled settings, case studies enable a detailed examination of interactions between technical, organizational, and human factors.

Moreover, Brocke et al. (2020) presents the design science research (DSR) strategy. This aims to address identified problems or phenomenon through the creation of artifacts. Oftentimes, the artifact is complemented by design knowledge providing relevant theory that support the artifact and its creation. According to Brocke et al. (2020), it is common to include other research methods when conducting a DSR study, such as interviews and literature reviews.

The DSR framework presented by Brocke et al. (2020) consists of three components: environment, design, and knowledge base. The environment component refers to the context in which the investigated phenomena takes place. Technology, people and organizations are all included in this component, providing an overview of the process setting. The knowledge base consists of existing knowledge and methodologies, guiding the research to reaching and evaluating a solution. The environment component ensures the solution's relevance, while the knowledge base provides rigor to the research. Finally, the design component combines the two, resulting in a solution-oriented process, using reliable methods to reach innovative solutions. The

framework is illustrated below in Figure 2.3.

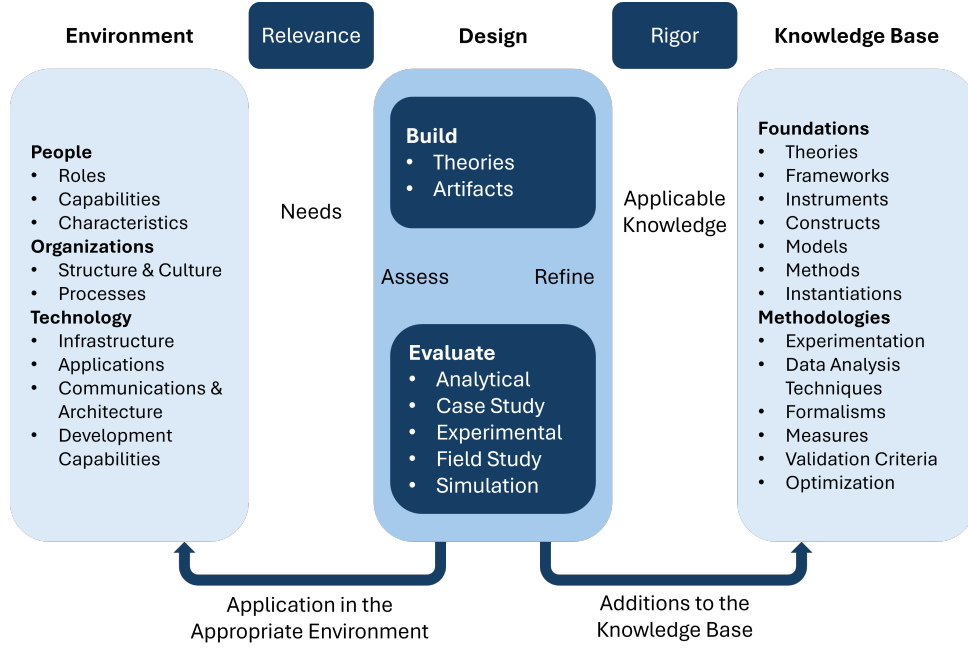


Figure 2.2: *DSR framework according to Brocke et al. (2020)*

When working with the framework, the most commonly adopted DSR process consists of six activities: problem identification and motivation, definition of the objectives for a solution, design and development, demonstration, evaluation, and communication (Brocke et al., 2020). The first activity defines the problem and why a solution is valuable. The second activity aims to describe how the research addresses the problem described in the problem identification and motivation step. In the third activity the artifact is created, and in the fourth activity this is applied to the problem, demonstrating its functionality. The fifth activity measures the performance of the artifact, and lastly, in the sixth activity, both the problem and the artifact are presented in their entirety.

This thesis adopts a DSR strategy, in which a case study is used to define and address the research phenomenon. The artifact to be developed is a prompting framework designed for data analysis and visualization. The empirical investigation is limited to a single company, making the study a single case study. Although this limits statistical generalizability, it allows for a contextualized understanding of AI-supported decision-making within a digital supply chain environment, where technical, organizational, and user-related factors are deeply intertwined (Höst et al., 2006). The problem identification has already been made in 1 and step two is addressed through this chapter.

PC SCM serves as the focal case due to its role as a provider of digital supply chain solutions and its access to operational data. Thus, this pose as the environment part of the DSR framework. By studying this single case in detail, while drawing on data from multiple sources, the thesis aims to generate analytical insights into both the

opportunities and limitations of AI-driven decision support in operational supply chain contexts. While the findings are not intended to be statistically generalizable, they provide analytical insights that may be transferable to similar digital supply chain environments and highlight areas that will require further exploration.

2.2.1 The Case Study

Denscombe (2010) defines a case study approach as a method that focuses on a limited number of instances in order to develop a detailed understanding of underlying events and processes. According to Yin (2018), case studies are particularly suitable when “how” questions are posed, the researcher has limited control over behavioral events, and the focus is on a contemporary phenomenon. In this thesis, the contemporary phenomenon under investigation is the integration of AI into digital supply chains. As previously stated, the case study centers on PC SCM, a Swedish software as a service provider specializing in digital supply chain networks. The study investigates how PC SCM’s existing platform can be enhanced through AI-driven KPI interpretation and decision-support.

Yin (2018) distinguishes between four types of case study designs based on two methods: single versus multiple cases, and holistic versus embedded units of analysis. Furthermore, Voss et al. (2002) identify four types of case studies: exploratory, theory building, theory testing, and theory extensive, each differing in purpose and structure. Moreover, according to Yin (2018), six main sources of evidence can be used in case study research: documentation, archival records, interviews, direct observations, participant observation, and physical artifacts. Employing multiple sources of evidence is recommended, as this enables triangulation and provides a more comprehensive understanding of the case.

The thesis adopts a combination of an exploratory and a theory testing approach. This is done by evaluating theories related to prompting genAI, and more specifically through a screening of existing frameworks, which is then applied to data from the case environment, with the objective of improving the user experience. Although the study focuses on a single focal company, it examines multiple units of analysis, such as different KPI categories and data from different industries, thereby adopting a single embedded case study design as illustrated in Figure 2.3. This design was selected primarily due to the time constraints of the thesis, while still allowing for analytical depth. The thesis relies primarily on archival records that provide the data used to test the AI-model. In addition, interviews are conducted to gain insights into critical factors relevant to the genAI analysis and to validate and contextualize findings derived from the archival data.

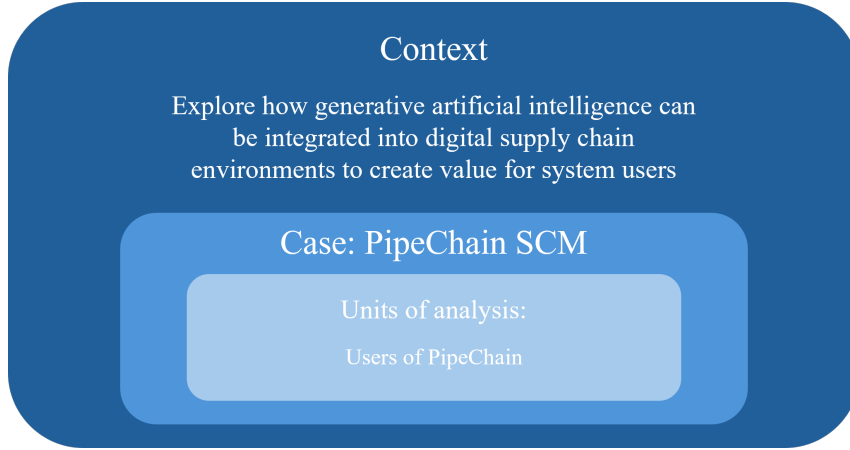


Figure 2.3: *The single-embedded case study design adopted in this study, developed by Yin (2018).*

2.3 Data Collection

Data collection in this thesis is conducted using a combination of qualitative and quantitative methods in order to obtain a comprehensive understanding of the system's requirements and performance. This approach constitutes a mixed methods research design, as defined by Schoonenboom (2023). While mixed research may present challenges, particularly due to the lack of universally accepted definitions of qualitative and quantitative data, a clear distinction between the two data types is essential for successful implementation. When applied appropriately, a mixed model research enables mitigation of the limitations associated with relying on a single type of data. In this thesis, data expressed numerically are classified as quantitative data, whereas data expressed in words are classified as qualitative.

2.3.1 Literature Review and Framework Screening

A literature review serves as an initial step to establish the theoretical foundation of the thesis and to support and validate its findings. The literature review consists of the collection and analysis of secondary data, with the primary objective of building upon existing research to ensure a solid scientific foundation, and to identify gaps further supporting the works of this thesis. This review is descriptive in nature and follows an iterative three-step process in accordance with Höst et al. (2006):

- *Broad search*
- *Selection*
- *Deep Search*

The broad search involves the use of a wide range of keywords to explore numerous research paths and obtain a comprehensive overview of the subject area. Based on an initial review of the material identified in the first stage, the most relevant sources are selected for further analysis. In the deep-search phase, specific terminology identified in earlier stages is used to explore reference lists of key publications and

to identify influential authors within the field. The literature review is structured around two main thematic areas:

(1) Supply Chain Management

This section introduces the context of the thesis by providing an overview of fundamental SCM concepts, key functional areas, and recent trends. It also establishes the theoretical foundation for performance measurement in supply chains, covering relevant frameworks, and the role of KPIs in decision-making.

(2) Artificial Intelligence

This section introduces AI as a concept, along with selected sub-fields relevant to the thesis, to provide a foundational understanding of the technologies. It further outlines two primary approaches to the use of AI, as well as general considerations regarding reliability. In addition, the section addresses the evaluation of AI-generated outputs, thereby establishing a foundation for the practical component of the thesis.

The chapter concludes with a third section that integrates the two areas by examining the application of AI within the SCM context. This section explores how AI is currently utilized in practice and serves to motivate the relevance of the research questions, position the thesis in relation to existing literature, and identify similarities and differences between the findings of this study and prior research. In line with Rowley and Slack (2004), the literature review also aims to identify research gaps, clarify key concepts and terminology, and provide a foundation for the analysis and interpretation of the empirical findings. Thus, in addition to 1, this sections serves as part of the first step of the DSR framework by providing a clear motivation of the study.

The literature search process follows the recommendations of Rowley and Slack (2004) and is conducted primarily using the university library database Finn. Advanced search functions were used, including keyword combinations, truncation, boolean operators, and filtering. Following the completion of the deep search, the literature review is developed through systematic scanning of sources, note-taking, structuring of content, and compilation of the bibliography.

A literature screening is also conducted when establishing a framework for prompting, following the same principles as the literature study. For this, in addition to Finn, the database Scopus is also consulted. This literature screening constitutes the third step of the DSR framework: design and development and is further expanded on in Chapter 5. The combination of the literature review and the framework screening constitutes the knowledge base in the DSR framework.

2.3.2 Quantitative Data

According to Sreejesh et al. (2013), quantitative data can be categorized as either primary or secondary data. Primary data is collected directly by the researcher for a specific research purpose, whereas secondary data have been collected previously for a different purpose but can still provide valuable insights. An advantage of using

secondary data is that it requires significantly less time to access, which in turn enables the use of a large sample size in the analysis.

Regarding quantitative data, this thesis relies on secondary data obtained from PC SCM's system, consisting primarily of anonymized historical supply chain data from the case-specific process. This data serves as input for evaluating genAI's ability to interpret changes in KPIs and identify potential risks. PC SCM's system collects and stores data daily, and visualizes it in either a daily or weekly format. The required datasets will be accessed through export from the PC platform to Microsoft Excel. Additionally, part of the quantitative data collection of this thesis consists of primary data generated from expert assessment of genAI output.

2.3.3 Qualitative Data

A common approach to collecting qualitative data in case study research is through interviews (Höst et al., 2006). These can be either of a structured, semi-structured, or unstructured character, where the primary distinction lies in the degree of pre-defined questions. Structured interviews are conducted by strictly following a pre-defined set of questions. The purpose of semi-structured interviews is to gain an in-depth understanding of the interviewee's qualitative and quantitative experiences, thereby enabling both description and explanation of the phenomenon studied. Unstructured interviews mostly lets the interviewee decide what to discuss.

To ensure the collection of reliable and relevant qualitative data, a stratified selection of interviewees is applied. According to Höst et al. (2006), stratification is a sampling strategy that is used to capture variation within a defined population rather than to achieve statistical representation. This involves first defining a relevant category of respondents and then selecting individuals within that category. In this thesis, the interview category is defined as individuals with substantial expertise in PC SCM's customers and system usage. Hence, internal system consultants and system users from PC SCM's customers are chosen as interviewees. This stratification enables the study to capture perspectives from both system providers and end users, ensuring a comprehensive understanding of the decision-making needs and system integration.

In this thesis, semi-structured, focused interviews are conducted with seven key employees at PC SCM as well as one client representative. The time scheduled for each interview is 20 minutes, during which one researcher is asking the questions and another is taking notes. However, the interviews are not limited to 20 minutes. Due to the presence of two researchers during the interviews, the possibility of seeking clarification from interviewees after the interviews, and the need to comply with the signed non-disclosure agreement, the interviews were not recorded. The interviews aim to map current decision-making requirements and challenges related to how users interact with, and interpret KPI data. Additionally, interviews serve as a source for what areas genAI may be able to create value. In addition to interviews, qualitative data is also gathered through collection generated text from the output provided from the genAI models.

2.4 Data Analysis

Given the mixed research approach, both qualitative and quantitative data are analyzed in order to address the research questions. According to Mays and Pope (2000), qualitative and quantitative data require different analytical methods to be examined efficiently. Consequently, each data type is analyzed using methods appropriate to its characteristics.

2.4.1 Quantitative Data

According to Höst et al. (2006), quantitative data can be used either to understand the phenomenon studied or to present connections and hypotheses that have been previously stated. In this thesis, the quantitative evaluation is conducted by focusing on two main areas with the objective to better understand how genAI can be used to analyze and present insights from supply chain data. Firstly, computational correctness and reliability is evaluated, by comparing the output of the AI model with manual calculations. This manual check also aims to identify issues that the AI potentially failed to find. Secondly robustness of the analysis of the genAI is examined through identical iterations and cross-model testing. Furthermore, the qualitative data generated by AI and assessed by experts, is scored on a number of aspects and analyzed quantitatively. This follows the methodology adopted by Yeen et al. (2024), by scoring each analyzed aspect on a scale from one to five.

2.4.2 Qualitative Data

According to Höst et al. (2006), qualitative analysis fundamentally concerns structuring the collected data in a systematic manner. Quantitative data can be processed using statistical techniques, but qualitative data require analytical methods based on sorting, categorization, and interpretation. Höst et al. (2006) presents four main approaches to qualitative analysis:

- *Quasi-static methods*
- *Template-based methods*
- *Editorial methods*
- *Deepening methods*

Quasi-static methods compare the frequency of key words and concepts to identify which topics are most prominent among respondents. Template-based methods rely on existing literature to define a predefined set of key concepts and focus on who mentions a topic rather than how frequently it appears. Editorial methods also categorize qualitative data but do not rely on predefined concepts; instead, categories emerge from the data itself. Finally, deepening methods are based on the researcher’s creativity and intuition when interpreting the data. However, due to the difficulty of systematically describing this process, Höst et al. (2006) argue that such methods cannot be considered fully scientific.

In addition, Mezmir (2020) presents a four-step framework for analyzing interview data:

- *Data familiarization*
- *Identifying a thematic framework*
- *Data reduction*
- *Data coding*

Data familiarization involves reviewing the collected data to gain an understanding of recurring themes, responses, and key issues. The second step focuses on developing a structure for organizing the data, such as categorizing responses and identifying commonly mentioned risks, as well as defining how the coded data will be presented. The third step, data reduction, entails filtering out data that are outside the scope of the study or do not contribute to answering the research questions. Lastly, data coding involves the systematic categorization of the remaining data to enable interpretation and the derivation of conclusions. Following initial coding, axial coding is applied to identify relationships between categories and establish key insights.

This thesis follows the framework presented by Mezmir (2020) while adopting a quasi-static approach, with elements of an editorial approach. The purpose of collecting qualitative data is to identify what topics and KPIs are considered the most important or are monitored most frequently when working within the PC SCM system. A quasi-static method is therefore well suited for initiating an analysis of patterns of convergence and divergence across interview responses and the literature study. The editorial approach then allows for categorization of areas of interest. The qualitative findings will then act as a guide for what aspects the AI model should primarily focus on when handed data.

2.5 Research Quality

To ensure overall research quality, three key aspects are commonly considered: reliability, validity, and representativeness (Höst et al., 2006). Reliability concerns the consistency and trustworthiness of the data and research procedures, validity refers to the extent to which the study measures what it intends to measure, and representativeness relates to the generalizability of the findings. As this thesis is conducted as a case study, representativeness in the statistical sense is not the primary objective. Nevertheless, a detailed description of the research context is provided to support the analytical generalization and transparency.

Reliability is addressed by ensuring accuracy and consistency throughout data collection and analysis. This is achieved through clearly defined procedures, systematic documentation of the research process, and continuous supervision during the study. In addition, the use of semi-structured interview protocols and reproducible data extraction procedures contributes to increasing the reliability of the study.

An important aspect of validity concerns the credibility of the literature used. To assess source quality and reduce the risk of bias, the literature review follows the

evaluation principles proposed by Höst et al. (2006):

- Is the material reviewed, and if so by who?
- Who guaranties the credibility of the source?
- Is the research methodology credible?
- Is the source in a context that is relevant for my research questions?
- Are the results supported or confirmed by other research, and have they been referred to in other credible contexts?

The primary objective is to base the thesis on reliable and well-supported sources.

Moreover, the validity of the presented prompting framework is addressed through empirical testing using the SAFE-PE framework presented by Laeeq and Saleem (2025), and further expanded on in Chapter 6. Test results are analyzed to understand strengths and weaknesses of the framework, thereby identifying areas where it might be further developed to achieve the desired output from genAI.

Finally, though the use of data from multiple companies and industries provided by PC SCM, the study increases the robustness of its findings and enables the evaluation of AI performance in a variety of supply chain contexts. Confidentiality is addressed through data anonymization and comprehensive documentation of the research process, thereby ensuring objectivity and integrity throughout the study. As the thesis is highly centered on a real supply chain process, the evaluation of relevant system characteristics is critical to ensuring high research quality. In this context, particular emphasis is placed on characteristics related to decision-making support.

2.6 Summary of Methodology

In summary, the thesis uses a DSR approach to create a prompting framework and test this in a real setting. Initially, a literature study and an interview guide is constructed, and then followed by the semi-structured interviews. These interviews create the foundation for the mapping of the case process which in turn provides insights to consider when creating a prompting framework. This after creating a framework this is adjusted and later tested and evaluated. Throughout this process, as Figure 2.4 illustrates, the abductive approach of alternating between theory and empiric is adopted.

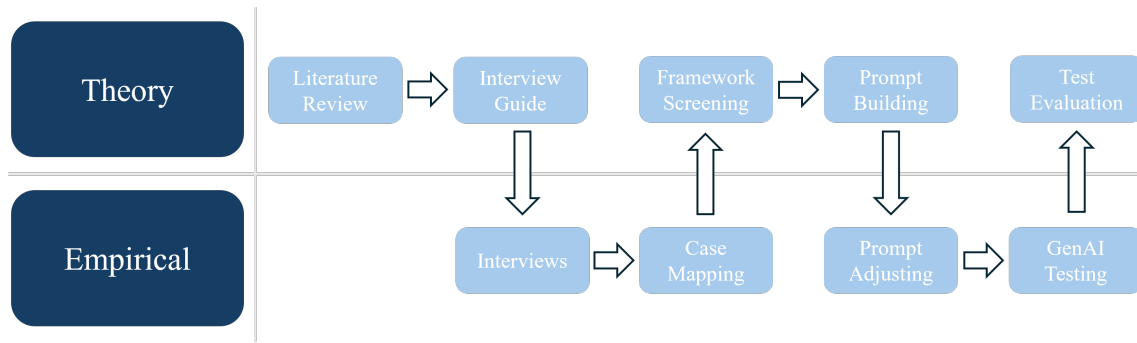


Figure 2.4: *An overview of the abductive research approach of the thesis*

3 Literature Review

This chapter is centered around two main areas; supply chain management and artificial intelligence. Each section aims to offer an understanding of the theoretical foundation of the thesis, and to support and later on validate the findings of the thesis. It will end with a combination of these two areas, discussing the use of artificial intelligence within the supply chain management context, after which the chapter is summarized and the relevance of the thesis is clarified from a literature perspective.

The literature review is compiled with the objective of providing foundational knowledge and a basis for analysis. As previously stated in Chapter 2, this chapter is divided into two main areas that together contribute to a general understanding of how AI may be used to further enhance PC SCM's offering within SCM. As presented in Figure 3.1 below, the study is centered on the areas *supply chain management* and *artificial intelligence*. The intersection of these areas presents the field of analysis for this thesis and will also be discussed in this chapter.

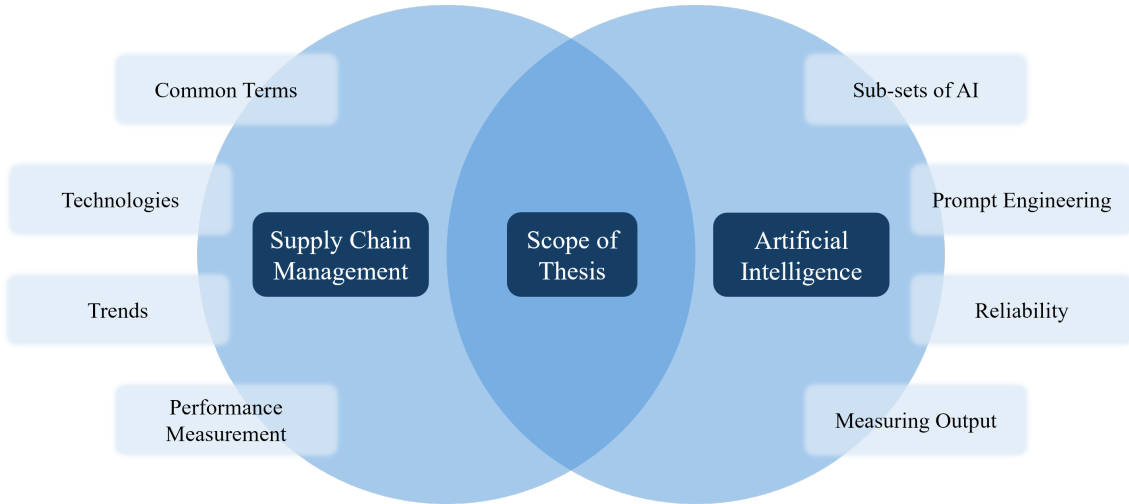


Figure 3.1: *Visual representation of the theoretical framework and its key components*

3.1 Supply Chain Management

Although precise definitions of the term supply chain tend to vary between researchers, most definitions align closely with each other. The term has evolved over time and innovations have affected what is considered part of SCM (Min et al., 2019). Olhager (2019) defines a supply chain as “a system of collaborative units which plan, control and develop the flow of material and information to deliver products and services from raw materials to final product”. He then continues to state that a company’s supply chain consists of three main functions; purchasing, production and distribution. Supply chains also contain material and information flows,

with material flows mainly flowing downstream and information flows often moving upstream. The conceptual supply chain presented by Olhager (2019) is visualized in Figure 3.2.

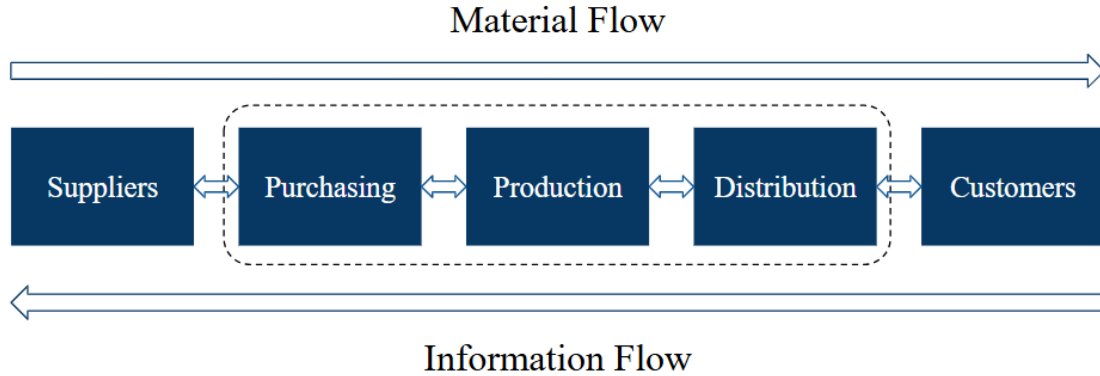


Figure 3.2: *A conceptual supply chain according to Olhager (2019)*

Purchasing

Today, purchasing plays a crucial role in the overall performance of companies and it enables value creation internally by acquiring materials, products, and services (Di Mauro et al., 2024). The purchasing process itself, however, creates costs for companies, and reducing these costs by establishing a well-performing supplier network, where the appropriate suppliers are chosen for the right products, contributes to the strategic goals of companies as well as strong revenues (Olhager, 2019).

Purchasing can be divided into strategic and operational purchasing. Strategic purchasing consists of working with long term issues, such as supplier selection, supplier development, and supply chain risk management (Olhager, 2019). It also includes several internal business units that work cross-functionally to collaborate. A stated success factor in achieving a well-performing supply chain is to extend this collaboration across company borders and thus increase transparency, collaboration, and traceability in the chain (Holmberg, 2000). This applies both upstream to suppliers and downstream to customers (Olhager, 2019). The handling and distribution of information can be as important as the orchestration of the physical material flow in purchasing. Operational purchasing activities include releasing purchase orders, monitoring supplier performance, and general work with the replenishment process (Maslaric et al., 2020).

Regarding SCM in general, and purchasing in particular, having well established collaboration with suppliers is seen as a competitive advantage compared to working alone. An example of a collaborative initiative used in industry is collaborative planning, forecasting, and replenishment, which was developed in the mid-90's. The idea behind it is to jointly forecast demand and plan production across company borders, enabling more optimized planning and reducing the risk of stock-outs. (Hollmann et al., 2015)

Vendor Managed Inventory

A strategy many companies, especially larger organizations, have implemented in their supply chain is VMI. VMI refers to a company outsourcing its inventory management of a component or product to the supplier of that product. Often, the buyer gives its VMI suppliers upper and lower stock level limits, between which the supplier is supposed to maintain the stock level through replenishment (Van den Bogaert and Van Jaarsveld, 2022; Olhager, 2019).

Compared to the buying company managing inventory on their own, VMI has been shown to enable several benefits for both suppliers and buyers (Van den Bogaert and Van Jaarsveld, 2022). The shared stock between the supply chain members enables a lower total stock level since the uncertainty is decreased through the information sharing between the parties. This entails less capital tied up in warehouses as well as a shorter throughput time for components. Additionally, VMI enables shorter response times for the supplier and thus increases flexibility in the supplier's production facility (Olhager, 2019).

Information Sharing and Collaboration

Having access to relevant information improves corporate decision-making at the strategic, tactical, and operational levels (Kache and Seuring, 2017; Kembro et al., 2017). With modern information systems that give users access to more data, companies are now facing a new challenge in prioritizing what information is the most relevant for their supply chain operations (Kache and Seuring, 2017). Having a thorough approach when creating a performance measurement system (PMS) is crucial, as supply chain performance is heavily influenced by information.

Building on this, data reliability is also of great importance in regards to decision-making in supply chain contexts (Kache and Seuring, 2017). Not having access to up-to-date and relevant data will reduce the quality of decision-making, and thus risk affecting the overall supply chain performance.

However, sufficient information should not be kept in-house, but rather be shared between supply chain members, inter-organizationally, to increase visibility and transparency throughout the supply chain (Kache and Seuring, 2017). Without proper information sharing, conflicts and misunderstandings may occur between supply chain members, reducing the ability of firms to compete. Being able to better take advantage of market opportunities, introduce new products, and problem solving are all activities that become more efficient in collaborative supply chain environments according to Baah et al. (2022).

Moreover, sharing information builds trust as supply chain partners feel integrated and updated when participating in a broader information flow (Baah et al., 2022). Consequently, supply chain members feel encouraged to work as a single unit, creating resilience to uncertainties and risks.

However, whilst many researchers emphasize the importance and benefits of information sharing across the supply chains, Kembro et al. (2017) highlights the challenges associated with implementing it in practice. Particularly when supply chains become more complex and consists of multiple tiers, companies often find it

increasingly difficult to realize the benefits of information sharing.

Although there is substantial research regarding information sharing in supply chains, much of it focuses on dyadic supplier-buyer relationships rather than multi-tier supply chains (Kembro et al., 2017; Kembro et al., 2025). Consequently, empirical evidence suggests that the implementation of information-sharing systems across entire supply chains remains limited (Kembro et al., 2025).

To explain these challenges, Kembro et al. (2017) identifies six categories of antecedents to multi-tier information sharing:

- information utilization
- technology utilization
- power structure
- culture
- business process
- legal aspects

These factors are critical for successful information sharing but can also act as significant barriers when not properly addressed.

Furthermore, Kembro et al. (2025) introduces the importance-feasibility paradox. As supply chains become more complex, the need for information sharing increases, yet the feasibility of implementing and maintaining effective information-sharing practices decreases. As a result, many organizations struggle to achieve the level of supply chain-wide information sharing that is often advocated in the literature.

Technologies

Today, a major part of the success of SCM is the ability to take advantage of the large amount of information available to companies. The general trend in supply chains is to become increasingly interconnected, relying heavily on technology rather than geography only (Rasool et al., 2023). Kache and Seuring (2017) considered the incorporation of big data analytics (BDA) a “management revolution” and a key driver to capture business value. Companies do this through the use of information systems that enable automated data handling and sharing between coworkers and across company boundaries. Several information systems are available to companies, and generally more than one system is used within one organization. Often companies have one enterprise resource planning system as a foundation, which is then complemented by process-specific systems. (Olhager, 2019)

A process specific system companies might use is a supplier relationship management (SRM) system. The purpose of an SRM system is to manage and develop partnerships between businesses and their suppliers (Emon et al., 2024; Olhager, 2019). Today, SRM is related to support of strategic decision making rather than transactional interactions, and encompasses risk management, supplier development, collaborative partnerships, and sustainability (Emon et al., 2024). An SRM system

enables companies to identify supplier relationships that can be improved, which relationships have poor communication performance, and where supply-related risks need to be managed.

3.1.1 Supply Chain Trends

Supply chains are undergoing significant transformation, driven by increasing global uncertainty and rapid technological development (Ghobakhloo et al., 2025; Huang et al., 2023). In recent years, both external disruptions and advancements in digital technologies have reshaped how organizations design, manage, and evaluate their supply chain operations (Ivanov, 2025; Singh, 2025; Ghobakhloo et al., 2025). Two prominent trends have emerged in this context: the growing emphasis on supply chain resilience and the adoption of Industry 4.0 technologies (Singh, 2025; Ghobakhloo et al., 2025). Together, these trends reflect a shift from efficiency-driven supply chains toward more adaptive, data-driven, and robust systems capable of responding to an increasingly complex and dynamic environment (Ghobakhloo et al., 2025; Huang et al., 2023).

Supply Chain Resilience

The COVID-19 pandemic highlighted issues in many supply chains (Ghobakhloo et al., 2025; Singh, 2025; Ivanov, 2025), exposing their vulnerability to disruptions. Since then, companies have changed their mindset towards “disasters will happen” and are configuring their supplier setup to be able to respond. Risks might include government regulations, political sanctions, and wars that begin near suppliers. Hence, the role of purchasing is becoming more critical for companies (Di Mauro et al., 2024).

The concept of supply chain resilience has become an area of attention for companies in the aftermath of the pandemic. Singh (2025) defines a resilient supply chain as “the ability of the supply network to remain operational despite disturbances” which is also connected to the ability of companies to adapt and improve performance when faced with disruptions in their supply chain (Ghobakhloo et al., 2025; Huang et al., 2023). Disruptions are a major vulnerability for supply chains and expose organizations to uncertainty and variability in demand and resilience requires companies to work collaboratively with their supply chain partners (Singh, 2025).

Supply chain resilience in companies is built through investment (Ivanov, 2025; Wu and Lam, 2026). Hence, a challenge when working with resilience is to evaluate what investments are worth doing from economic and sustainable perspectives. Furthermore, building resilience adds value long-term, but tends to be assessed short-term. According to (Wu and Lam, 2026), this approach does not reflect reality, as resilience investments add value when disrupting events actually occur. Short-term impact measurements are likely to miss these events and therefore underestimate the impact of resilience on supply chain performance.

Supply chain resilience as a concept also existed before the pandemic. Tukamuhabwa et al. (2015) conducted a literature review on the topic in 2015, where many findings

align well with current research. Political turmoil, disease, terrorism, and natural disasters were all discussed as potential threats to supply chain operations. Furthermore, the authors divide strategies for building supply chain resilience into proactive and reactive strategies, but also state that some strategies can be used in both categories. An example of this is supply chain collaboration, which can build trust and security proactively, as well as enabling coordinated responses to disruptions (Tukamuhabwa et al., 2015).

Industry 4.0 Supply Chains

During the last decade, the industry has gone through what researchers call the fourth technological revolution (Ghobakhloo et al., 2025; Huang et al., 2023). The new age of technology is known as Industry 4.0, and several technological breakthroughs are considered to be part of this new age. AI, big data analytics, blockchain, and the internet of things are examples of technologies related to Industry 4.0. Integrating these into supply chain processes enables greater resilience (Ghobakhloo et al., 2025). Huang et al. (2023) identify 16 technologies that are considered part of Industry 4.0:

- Artificial intelligence
- Augmented reality
- Autonomous vehicle
- Blockchain
- Big Data Analytics
- Business Intelligence
- Cloud computing
- Enterprise resource planning
- Internet of things
- Machine to machine
- Radio frequency identification
- Robotics
- Sensors
- Simulation
- Virtual reality
- 3D printing

Ghobakhloo et al. (2025) identifies a variety of benefits that supply chains receive from Industry 4.0. In terms of achieving supply chain resilience, the technologies offer improved real-time data access, enabling the traceability of current operations and the identification of future disruptions. Furthermore, BDA allows for improved

performance measurements in all business units at the strategic, tactical, and operational levels (Ghobakhloo et al., 2025). The planning process has become more data-driven since the introduction of Industry 4.0 (Bueno et al., 2020).

Information flows are being created more effortlessly with 4.0 technologies. BDA and cloud-based software allow easier access to valuable information, and thereby reduce the risks of information silos being created between business units or company borders. Consequently, Industry 4.0 technologies promote transparency and trust within supply chains (Ghobakhloo et al., 2025; Huang et al., 2023).

Furthermore, the introduction of Industry 4.0 technologies has allowed companies to automate their business processes to a greater extent. Flows of information, finances, and physical materials require less manual input than before, allowing a lower dependence on human intervention (Ghobakhloo et al., 2025).

To complement the focus of Industry 4.0 on efficiency through automation and integration of technology, European Commission (2021) coined the term “industry 5.0”, which focuses on how technologies can be used in a sustainable, human-centric, and resilient way. While Industry 4.0 focuses on automation and technology, industry 5.0 refers to how these technologies can support employees (Fares et al., 2025). In addition, resilience is an important focus when companies look to integrate industry 5.0 into their supply chain operations. This is part of the COVID-19 aftermath, where handling disruptions effectively has become a necessity for companies.

3.1.2 Performance Measurement

Performance measurement can be defined as the process of quantifying the efficiency and effectiveness of organizational actions (Shepherd and Günter, 2006; Gunasekaran and Kobu, 2007). It is operationalized through PMS’, which consist of the metrics used to evaluate organizational performance (Shepherd and Günter, 2006; Marinagi et al., 2023). The primary purpose of performance measurement is to support organizational decision-making by providing objective and relevant information. (Holmberg, 2000).

Performance measurement serves several important organizational functions. It helps identify success factors, performance gaps, bottlenecks, and waste, while also enabling organizations to assess whether customer needs are being met (Lakri and Jemai, 2015; Gunasekaran and Kobu, 2007). In addition, PMS support the communication of organizational strategy by translating strategic objectives into measurable targets and facilitating organizational learning through feedback, performance evaluation, and continuous improvement (Lakri and Jemai, 2015).

Historically, PMS relied primarily on financial indicators such as profit margins and cash flows, which mainly reflected the outcomes of past actions (Nazari-Ghanbarloo, 2020). However, modern performance measurement increasingly incorporates non-financial and intangible indicators, supported by digital technologies that enable real-time monitoring and early detection of operational disruptions (Holmberg, 2000; Gunasekaran and Kobu, 2007; Ponce et al., 2025; Marinagi et al., 2023; Lakri and Jemai, 2015). This development reflects the growing complexity and dynamic nature

of organizational environments.

In supply chain contexts, performance measurement is particularly important due to the interdependence between organizations and processes. As competition increasingly occurs between entire supply chains rather than individual firms, effective performance measurement has become essential for coordination and strategic alignment across the supply chain (Varadejsatitwong et al., 2022). In this context, supply chain performance measurement enables organizations to evaluate operation, identify inefficiencies and bottlenecks, monitor progress, benchmark against other organizations, and support informed decision-making and continuous improvement (Buchari and Tontowi, 2024; Balfaqih et al., 2016; Lakri and Jemai, 2015; Rasool et al., 2023).

An important prerequisite for effective measurement of supply chain performance is that supply chain partners establish shared measurement systems that provide relevant and comparable information (Holmberg, 2000). However, organizations often place an excessive emphasis on financial indicators despite the fact that supply chain performance is characterized by complex and dynamic processes. Although financial metrics remain important for monitoring economic results, they are less effective in improving operational activities and coordination across supply chain actors (Holmberg, 2000).

In recent years, digitalization has significantly increased the availability of operational data within supply chains. Digital supply chain systems generate large volumes of data that allow organizations to measure performance in new ways (Özbiltekin-Pala, 2026). As a result, digital performance measurement has become increasingly important for monitoring supply chain operations and identifying potential disruptions. This trend has accelerated following the COVID-19 pandemic, which highlighted the importance of resilience and proactive risk management within the supply chain (Özbiltekin-Pala, 2026). Enhanced visibility through performance measurement allows companies to access relevant information more efficiently, while low visibility can limit supply chain resilience and increase exposure to disruptions and stock-outs (Phung et al., 2024; Ghobakhloo et al., 2025).

Several frameworks have been developed to structure and guide performance measurement in organizational contexts. One widely recognized example is the Balanced Scorecard, which evaluates organizational performance across four perspectives: financial, customer, internal processes, and learning and growth (Varadejsatitwong et al., 2022). Another widely used framework in supply chain contexts is the Supply Chain Operations Reference (SCOR) model, developed by the Supply Chain Council. The model offers methodology, diagnostic tools, and benchmarking capabilities that allow organizations to evaluate and improve their supply chain performance (Association for Supply Chain Management, 2022; Harmon, 2003). As a process reference model, SCOR defines supply chain processes in a way that aligns them with key business functions and strategic goals. Despite being introduced back in 1996, the SCOR model remains widely used for benchmarking supply chain performance across industries. The framework is built around four main components:

- Performance, which defines strategic objectives and standardized metrics used

to evaluate process performance.

- Processes, which describe supply chain activities and relationships.
- Practices, which identify management practices that lead to superior performance.
- People, which define the skills and competencies required to execute supply chain processes.

Key Performance Indicators

KPIs are specific and measurable values that are used to quantify the efficiency and effectiveness of organizational activities (Ponce et al., 2025; Shepherd and Günter, 2006; Lakri and Jemai, 2015). Within PMS, KPIs serve as reference points for evaluating performance and interpreting complex operational data (Saenz, 2021). By tracking, interpreting, and communicating performance metrics, KPIs support decision-making at different organizational levels and are therefore central to performance management in supply chain organizations (Marinagi et al., 2023; Holmberg, 2000). Metrics usually consist of three components: a performance measure, a target value, and the consequences associated with meeting or failing to meet that target (Lakri and Jemai, 2015).

KPIs enable organizations to monitor operational performance, highlight achievements, and improve understanding of key processes. Through systematic monitoring, KPIs facilitate the identification of operational problems such as bottlenecks, waste, or inefficiencies, ensuring that decisions are based on objective performance data rather than intuition (Lakri and Jemai, 2015; Rasool et al., 2023). Consequently, KPIs represent an important tool for improving and optimizing supply chain performance (Cai et al., 2009). Finally, KPIs influence organizational behavior by directing attention to critical issues and aligning actions with strategic objectives (Lakri and Jemai, 2015).

3.2 Artificial Intelligence

It is argued that the foundations of AI were laid by Alan Turing around 1948 (Sunday Grève, 2025). Although Turing himself never used the term artificial intelligence in his publications, his work was highly centered around whether machines could be taught to think like humans. This line of inquiry provided a basis for the continued development of what is now known as AI and ML. Following Turing’s contributions, several researchers have made significant advances in the field, with John McCarthy being the first to formally coin the term artificial intelligence, describing it as the ability of machines to use language, form abstractions and perform tasks traditionally associated with human intelligence (Ferreira et al., 2025).

In the last ten years, research regarding AI has drastically intensified as areas such as deep learning and big data have emerged (Bahoo et al., 2023). This is also the case when looking at AI in the context of SCM. According to Ferreira et al. (2025), the number of publications since 2020 has increased drastically. This sudden boost can

be explained by technological drivers such as increased computing power, maturation of cloud infrastructure, and the availability of massive datasets (Helo and Hao, 2021; Roy et al., 2024).

Despite its long existence, there is no universal consensus on the definition of AI. Instead, definitions tend to vary depending on the function and scope attributed. Numerous researchers have attempted to provide overarching definitions of the concept. Ferreira et al. (2025) defines AI as the use of computers to perform reasoning, pattern recognition, learning, and understanding behaviors derived from experience. In contrast, Dash et al. (2019) describe AI as a computer’s ability to independently solve problems for which it has not been explicitly programmed. Bahoo et al. (2023) emphasize the data interpretation aspect of AI, defining it as the ability of a system to correctly interpret external data, learn from it, and apply the knowledge to achieve specific goals. Similarly, Roy et al. (2024) define AI as “the overall ability of a machine to think or mimic human behavior” while analyzing data.

From an engineering perspective, the objective of AI is to create rational agents capable of perceiving their environment and acting to optimize a specific objective function (Helo and Hao, 2021). Although these definitions differ in emphasis, they share a common underlying understanding of AI. In this thesis, the definition presented by the European Parliament and Council (2024) is adopted, which states that AI is defined as follows:

“a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments”

The following parts provides a more detailed examination of selected AI sub-fields relevant to this thesis.

Machine Learning

ML is a subset of AI focused on developing algorithms and methods that enable computer systems to learn from data and experience to solve problems, rather than being explicitly programmed to do so (Goutier et al., 2019; Roy et al., 2024; Ferreira et al., 2025). It is defined as the discipline of creating algorithms that generalize patterns from sets of examples, allowing a system to make predictions or decisions. From a technical perspective, the core mechanisms involve the ability of a system to automatically adjust its internal parameters to align more effectively with observed data. (Ferreira et al., 2025)

Depending on the nature of the data and the learning objective, there are three main predictive approaches distinguished (Heinrich et al., 2021; Roy et al., 2024). In *supervised learning*, the dataset consists of labeled observations in which inputs are paired with corresponding outputs. The algorithm learns a mapping function that enables it to predict correct targets for new inputs, commonly applied in tasks such as classification or regression. *Unsupervised learning*, by contrast, involves analyzing

unlabeled data, where the algorithm identifies patterns, hidden structures, or subgroups using techniques such as clustering or dimensionality reduction. Finally, in *reinforcement learning*, the system learns through trial and error by interacting with its environment, receiving rewards for desired actions and penalties for undesired ones.

Generative Artificial Intelligence

GenAI is a specialized branch of AI that uses ML and extends beyond data classification to the creation of new original content by learning and replicating patterns found in existing datasets (Ronge et al., 2025). Unlike discriminative models, which are primarily designed for classification or prediction tasks, genAI is grounded in generative modeling, where statistical approaches approximate the underlying probability distribution of a dataset, enabling the synthesis of new data samples that resemble the original source material. (Ronge et al., 2025; Triguero et al., 2023) By learning the structural characteristics of the data, generative models can produce realistic and contextually coherent outputs rather than merely identifying patterns (Triguero et al., 2023).

The field of genAI is currently shaped by three key technological approaches (Ronge et al., 2025; Triguero et al., 2023). First, *generative adversarial networks* consist of two competing neural networks that interact to produce highly realistic synthetic data, such as images. Second, *diffusion models* generate content by progressively removing noise from randomly initialized data, ultimately producing high-fidelity outputs. Third, *large language models* are foundation models trained on extensive text collections to learn linguistic probability distributions, enabling the generation of coherent and contextually relevant text. LLMs are often considered the primary form of genAI because they learn the probability distribution over their vocabulary to synthesize new coherent data instances (Triguero et al., 2023).

Large Language Model

LLMs are AI systems trained on vast and diverse datasets, enabling them to understand, process, and generate human-like text and multi modal content (Wang et al., 2025). Based on transformer architectures that employ encoder-decoder mechanisms, these models capture complex linguistic patterns and relationships, allowing them to perform a wide range of tasks, including pattern recognition, regularity identification, and text generation (Venkatachalam, 2025; Pang et al., 2024). LLMs are typically developed as foundation models, meaning that they are trained on large-scale, heterogeneous data using self-supervised learning, often followed by fine-tuning and reinforcement learning from human feedback. (Triguero et al., 2023) Examples of LLMs are chat bots such as ChatGPT, Gemini, and Copilot.

To generate and interpret human-like language, LLMs rely on natural language processing (NLP), a subfield of AI that enables machines to understand and produce human language in a meaningful manner (Bahoo et al., 2023; Helo and Hao, 2021). Traditionally, NLP has been defined as a discipline aimed at equipping computers with the capability to interpret their linguistic environment and act based on pat-

terns identified in text or speech. Methodologically, NLP approaches can be broadly categorized into *statistical NLP*, which relies on probabilistic analysis of words and phrases, and *semantic NLP*, which focuses on meaning representation through ontologies and structured linguistics. (Helo and Hao, 2021) NLP supports a wide range of applications, including chat bots and intelligent virtual assistants (Bahoo et al., 2023; Helo and Hao, 2021).

Within the supply chain context, NLP is mainly applied to extract meaningful information, such as entities, locations, topics, and sentiments, from large volumes of unstructured data (Helo and Hao, 2021; Ferreira et al., 2025). For more advanced analytical tasks, NLP functions as an interactive and explainable interface, translating complex analytical output into clear and accessible natural-language summaries (Venkatachalam, 2025).

3.2.1 Prompt Engineering

An input instruction provided to an AI model to direct its behavior and shape the resulting output is called a prompt (Giray, 2023). The field of creating such prompts, prompt engineering, is relatively new, having emerged in 2020, and refers to the process of designing and structuring inputs so that genAI systems can interpret instructions and produce desired output (Phillips-Wren and Håkansson, 2025). Several techniques and strategies for prompting have been developed (Vilakati, 2025), however, as highlighted by Phillips-Wren and Håkansson (2025), all prompts should be clearly formulated and include example tags that specify the desired structure of the output. The selection of technique to use is highly dependent on the desired output and the use case. A selection of techniques are further explained below.

Zero-Shot Prompting

The most elementary prompting technique when interacting with LLMs is zero-shot prompting, in which the LLM is given instructions without any examples (Laeq and Saleem, 2025; Kim et al., 2025; Phillips-Wren and Håkansson, 2025; Meyer et al., 2024). When using this technique, the model relies entirely on its pre-trained knowledge (Meyer et al., 2024), which can affect the reliability and relevance of the output (Kim et al., 2025; Phillips-Wren and Håkansson, 2025). The main advantages are flexibility, simplicity, and efficiency as it requires minimal effort and works well for straight-forward, well-defined tasks, thus not demanding much effort from the user (Phillips-Wren and Håkansson, 2025; Vilakati, 2025).

Few-Shot Prompting

Few-shot prompting builds on zero-shot prompting by providing the model with a small number of examples that demonstrate the expected output format or content (Laeq and Saleem, 2025; Kim et al., 2025; Vilakati, 2025; Phillips-Wren and Håkansson, 2025). In comparison to zero-shot prompting, this technique enhances both the reliability and the accuracy of the output and improves the performance through in-context learning (Kim et al., 2025; Phillips-Wren and Håkansson, 2025).

However, the quality of the output is heavily dependent and influenced by the examples provided, putting emphasis on the quality of these (Kim et al., 2025; Vilakati, 2025). The technique is also hindered by the number of examples provided as it is time consuming for the model to go through them, thus increasing the computational cost (Phillips-Wren and Håkansson, 2025)

Chain-of-Thought Prompting

The prompting technique chain-of-thought (CoT) encourages the model to explicitly articulate its reasoning process step by step before providing a final answer (Vilakati, 2025; Phillips-Wren and Håkansson, 2025; Kim et al., 2025). By breaking down the steps to solving problems, CoT improves transparency and explainability and addresses reliability issues, allowing the users to trace the sequence of decisions or calculations (Kim et al., 2025; Vilakati, 2025; Phillips-Wren and Håkansson, 2025). The technique is, however, not suitable for all problems and is prone to provide wordy output. Moreover, it does not necessarily guarantee that every step is correct as errors still can occur in calculations or logic (Kim et al., 2025; Vilakati, 2025).

Instruction-Based Prompting

Instruction-based prompting is a structured approach where the model is given clear, detailed, and step-by-step instructions to follow when completing a task (Vilakati, 2025). By outlining the sequence of actions, conditions, and checks, this method reduces ambiguity and guides the model through complex processes in a controlled way. It is particularly useful in tasks that require high precision, such as statistical or medical analysis, as it helps ensure that important steps are not overlooked, and it improves the reliability and transparency of the output. However, its effectiveness depends on the user’s ability to provide accurate and comprehensive instructions, which means that it requires a strong understanding of the task domain.

Summary

As presented, the different methods diverge regarding cost, complexity, and level of output control. In Table 3.1 the techniques presented above are summarized, along with their advantages and disadvantages.

Table 3.1: Comparison of prompting techniques

Technique	Description	Advantages	Disadvantages
Zero-shot	Instructions without examples, relying on pre-trained knowledge.	Simple, flexible, and efficient for well-defined tasks.	Lower reliability; depends entirely on existing model knowledge.
Few-shot	A few examples are provided to guide the expected output.	Improves reliability and accuracy through in-context learning.	Dependent on example quality; increases computational cost.
Chain-of-Thought	The model explains its reasoning step-by-step before answering.	Improves transparency and reasoning traceability.	Can produce verbose outputs and incorrect intermediate steps.
Instruction-Based	The model receives detailed and structured instructions.	Reduces ambiguity and improves precision in complex tasks.	Requires strong domain knowledge and precise instructions.

As the field of prompt engineering has developed and matured, hybrid strategies have evolved, combining different strategies with the objective to blend the different approaches’ strengths and give the user the necessary control of the output (Vilakati, 2025; Phillips-Wren and Håkansson, 2025). Vilakati (2025) suggests that a hybrid approach is the most advantageous as this often entails better robustness and reliability.

By using properly designed, high quality prompts, the user can safeguard against subtle errors or misinterpretations, and guide the AI towards reliable and relevant outputs. Regarding the design of prompts, there are four essential components that researchers agree should be included (Vilakati, 2025). These are:

- Clarity
- Context
- Instructions
- Guidance

Clarity emphasizes the importance of formulating a well-defined task or question within the prompt, ensuring specificity while avoiding redundant information that may distract from the objective. Context highlights the need to provide relevant background information to support the model’s understanding of the task. Instructions refer to explicitly specifying how the output should be generated, including

aspects such as format and style. Finally, guidance is particularly important for complex tasks, as it directs how the model should approach and solve the problem. Additionally, prompts should account for potential ambiguities, such as unclear terminology or multiple possible interpretations. (Vilakati, 2025)

3.2.2 Fine Tuning

In addition to prompt engineering, fine tuning is a technique that can be used with the objective to better specify the AI output. Fine tuning is a central technique in the development of LLMs, enabling pretrained models to be adapted to specific domains or tasks by further training on targeted datasets (Pang et al., 2024; Fatawi et al., 2024). Through this process, model parameters are adjusted to improve performance in a particular context while retaining the general knowledge acquired during pretraining (Fatawi et al., 2024). This capability is especially important as AI applications increasingly expand into specialized domains, where context-sensitive and accurate outputs are required. Prior research highlights that fine-tuning not only enhances task performance but also improves interpretability and relevance, for example in applications such as multivariate time series prediction and fault diagnosis in complex systems (Fatawi et al., 2024).

Instruction tuning represents a specialized form of supervised fine tuning, in which models are trained on structured instruction-output pairs (Pang et al., 2024; Çelikmasat et al., 2025). Unlike traditional fine-tuning that may focus on narrow input-output mappings, instruction tuning enables models to learn the underlying intent and structure of tasks, thereby improving their ability to generalize to unseen instructions (Çelikmasat et al., 2025). This makes instruction tuning particularly effective for downstream tasks requiring consistent structure and reasoning (Pang et al., 2024). Empirical studies demonstrate that instruction tuning significantly enhances zero-shot performance across a range of natural processing benchmarks (Çelikmasat et al., 2025).

In contrast to prompt engineering, which operates solely at the input level without modifying model parameters, both fine-tuning and instruction tuning embed task-specific behavior directly into the model (Çelikmasat et al., 2025). However, full-parameter fine-tuning is often computationally expensive, particularly for large-scale models (Lee and Park, 2025; Çelikmasat et al., 2025). Overall however, fine-tuning, and in particular instruction tuning, plays a crucial role in tailoring general-purpose LLMs to domain-specific applications. By balancing general knowledge with task-specific adaptation, and leveraging efficient training methods, these techniques enable the practical deployment of LLMs in complex and specialized use cases without the need for full model retraining.

3.2.3 Reliability

While the potential for AI in SCM is considerable, its implementation is often constrained by substantial technical and organizational challenges. Moreover, its adoption raises critical concerns regarding interpretability, trust, accountability, and organizational readiness (Aaijaz et al., 2025).

A central issue in the deployment of genAI is output reliability, particularly in relation to hallucinations (Rathkopf, 2026; Athaluri et al., 2023). Hallucinations occur when a model generates plausible yet factually incorrect or contextually irrelevant information that is not grounded in its training data (Triguero et al., 2023). Even highly advanced scientific systems have demonstrated such behavior, for instance the AI model AlphaFold 3, specialized in predicting molecular structures, has been reported to generate detailed molecular structures in incorrect locations (Abramson et al., 2024). These inaccuracies are difficult to systematically prevent or detect, thereby placing a significant burden on users’ domain expertise to validate AI-generated outputs and safeguard decision quality (Rathkopf, 2026). In this sense, hallucinations may resemble human speculation, particularly when the model operates under conditions of incomplete information (Triguero et al., 2023). However, in high-stake supply chain contexts, even minor inaccuracies may have considerable operational consequences.

Beyond reliability concerns, bias and sustainability represent additional areas of concern. Because AI models are trained on human-generated data, they may inherit and amplify existing societal or organizational biases, potentially leading to distorted or inequitable outputs (Wang et al., 2025; Huckenbeck and Zatrochová, 2024). Furthermore, the environmental impact of large-scale generative models is considerable, as their development and training require substantial computational resources and energy consumption (Triguero et al., 2023; Wang et al., 2025). These issues introduce ethical and strategic considerations that extend beyond purely technical performance.

Another significant challenge relates to the “black-box” nature of many AI systems (Bahroun et al., 2025). In complex models, it is often difficult to trace how specific outputs are generated, limiting transparency and interpretability. Castelvechi (2016) argues that even when an AI system produces accurate results, understanding the underlying reasoning process remains essential for ensuring accountability and trust. This lack of transparency raises important questions regarding organizational readiness, governance, and responsible deployment (Genetti et al., 2025; Huckenbeck and Zatrochová, 2024).

Finally, technical infrastructure and data management present practical barriers to effective implementation. The rapid expansion of internet of things devices and 5G connectivity has resulted in an unprecedented growth in data volumes, increasing the complexity of data governance and integration (Mulongo, 2024). Since AI performance is highly dependent on the quality, consistency, and integration of underlying data, deficiencies in data management can significantly undermine analytical reliability (Riad et al., 2024; Yerra, 2025). In addition, organizations must address cyber security risks and invest in workforce training to foster effective human-AI collaboration (Verma et al., 2025; Yerra, 2025). Ultimately, achieving trustworthy and balanced decision-making in supply chain environments requires not only advanced computational capabilities but also the integration of human judgment, contextual understanding, and organizational oversight (Riad et al., 2024).

3.2.4 Output Performance of Generative Artificial Intelligence

A challenge when investigating genAI's ability to perform tasks efficiently is to evaluate the quality of output data. According to Ramamoorthy (2025), evaluation approaches include both quantitative metrics and qualitative approaches. GenAI's production of quantitative data is often evaluated on reliability, for example if produced values are correct. Qualitative data on the other hand, tend to be adjusted based on research objectives and there is no standardized measurement model that fully captures the output quality from genAI (Ramamoorthy, 2025).

There are several examples of researchers evaluating output quality of GenAI. Yeen et al. (2024) conducted a study where ChatGPT was handed large data sets along with the task to analyze and present findings from the data. The study evaluated output by investigating three main areas in the GenAI output:

- Output Consistency
- Analytical Efficiency
- User Friendliness

Two researchers were handed output data and scored performance for each of the three areas individually on a 1-5 scale, where 1 is the lowest score and 5 the highest. The researchers also scored overall performance on a 1-3 scale, again where 1 is the lowest and 3 the highest.

Another example is (Laskar et al., 2023), who use independent, manual assessment to evaluate and compare output performance. The study also included correction of numerical values, while triangulation through at least two researchers looking at produced messages was used to evaluate qualitative outputs.

3.3 Artificial Intelligence Within the Supply Chain Management Context

The implementation of AI and ML is shifting organizations from reactive to proactive modes of operation (Yerra, 2025). By enabling real-time visibility, AI enhances decision-making and coordination, while improving transparency and accountability across the supply chain (Riad et al., 2024; Verma et al., 2025). Khokrale (2025) argues that AI-integrated supply chains extend beyond improved information processing capabilities; they enable synchronized responsiveness, resilience, and customer-centric performance. Furthermore, AI and ML play a crucial role in analyzing performance data and identifying patterns that would otherwise remain undetected (Mulongo, 2024).

Through their specialized capabilities, AI technologies significantly improve supply chain operations, with ML having a central role through predictive analytics (Riad et al., 2024). For instance, automated anomaly detection allows organizations to identify deviations from expected performance, thereby supporting smooth and reliable operations. By leveraging predictive analytics and historical data, organizations

can generate more accurate forecasts of future scenarios, including demand fluctuations, which in turn enhances inventory optimization. Empirical evidence suggests that AI-based predictive analytics strengthens responsiveness, planning accuracy, and logistics agility by delivering measurable end-to-end performance improvements (Khokrale, 2025; Verma et al., 2025). Moreover, the integration of predictive analytics, prescriptive optimization, ML, LLMs, and decision intelligence platforms enables enterprises to reduce decision cycle times, minimize errors, and optimize outcomes across strategic, tactical, and operational levels (Aaijaz et al., 2025).

Furthermore, the implementation of AI in SCM can reduce or even eliminate manual labor in certain areas by automating repetitive and rule-based tasks (Dash et al., 2019; Huckenbeck and Zatrochová, 2024; Verma et al., 2025). AI-supported decision-making systems transform large volumes of data into actionable insights, automate routine judgment processes, and augment human decision-makers with enhanced foresight and prescriptive guidance (Aaijaz et al., 2025).

More broadly, AI has become a key enabler of both planning and execution within supply chains. Core application areas include demand forecasting, inventory planning, transportation routing, and supplier risk management. (Khokrale, 2025; Verma et al., 2025). Regarding forecasting, planning (Dash et al., 2019; Verma et al., 2025), inventory management, and replenishment, predictive models help organizations determine optimal stock levels, locations, and timing, thereby supporting service level optimization and effective working capital management (Khokrale, 2025). By forecasting demand patterns more accurately, organizations can reduce the risk of shortages and excess inventory (Riad et al., 2024).

AI has also been widely applied in route optimization and scheduling, leading to significant reductions in time and cost (Riad et al., 2024; Verma et al., 2025). In addition, real-time data analytics and ML algorithms enable the prediction of shipping delays with high accuracy, allowing organizations to take proactive measures to mitigate disruptions and enhance operational efficiency (Yerra, 2025). Beyond logistics, AI creates value in production, promotion and pricing, delivery, and transportation (Dash et al., 2019). Automation and robotics, particularly in warehousing and production environments, have further transformed operations by increasing productivity and streamlining processes (Verma et al., 2025).

Across these application areas, AI contributes to substantial cost savings and provides practical solutions to a range of SCM challenges. As discussed, AI enhances responsiveness, agility, competitiveness, and operational effectiveness, enabling organizations to streamline processes, anticipate trends, and foster innovation (Verma et al., 2025). Aaijaz et al. (2025) demonstrate that AI-driven models outperform traditional approaches in terms of accuracy, operational efficiency, and adaptability. These findings confirm AI's role as a dynamic enabler of data-driven decision-making in modern supply chains.

Building on these developments, genAI represents further evolution of AI applications within SCM (Riad et al., 2024). While traditional AI and ML models primarily focus on prediction and optimization, genAI extends these capabilities by generating new data, scenarios, and solution alternatives. Its integration into SCM has

been associated with improvements in efficiency, transparency, sustainability, and adaptability. In practice, genAI enables organizations to move beyond retrospective analysis toward proactive and scenario-based planning (Ferreira et al., 2025). For example, models can generate synthetic supply chain scenarios, such as supplier disruptions or sudden demand fluctuations, to support decision-makers in evaluating alternative response strategies. Additionally, genAI can assist in developing and refining algorithms for demand forecasting, inventory planning, and network design, thereby strengthening decision support across strategic and operational levels (Dabbous et al., 2025).

3.3.1 Relevance and Research Gap

AI is increasingly being integrated into both everyday life and organizational processes, with companies across industries investing in identifying how AI can generate business value. In SCM, the growing availability of data, enabled by digital systems, has led to more detailed performance tracking than ever before. However, this development has also introduced new challenges related to data management, interpretation, and prioritization. While increased visibility across supply chains is essential for enhancing resilience, the abundance of available data, and thereby metrics, often requires a high level of user expertise. Without sufficient understanding of the underlying root causes behind these metrics, access to data may risk impairing, rather than improving, decision-making quality.

Existing research highlights the value of integrating AI into supply chain contexts, particularly due to its ability to process large datasets and identify patterns. AI applications have been shown to improve efficiency in areas such as forecasting, production scheduling, and transportation optimization. Furthermore, Industry 4.0 technologies enable dynamic and real-time performance measurement, allowing organizations to gain faster and more detailed insights into their operations. These developments suggest significant potential for AI to support decision-making by automating aspects of data handling and analysis.

Despite these advancements, current AI applications in SCM are predominantly focused on optimization and prediction tasks and there appears to be a lack of research exploring how genAI can be used to provide contextualized insights and explanations of data to support users in digital environments. In particular, limited attention has been given to how genAI can assist in interpreting performance data and identifying deviations, while simultaneously reducing the need for manual data filtering and analysis. This represents a critical gap, as misinterpretation of data remains a key challenge in data-driven decision-making processes.

However, recently a number of studies have been conducted to investigate genAI’s behavior in supply chain contexts. For example, Kirshner (2024) and Zhang and Cavusoglu (2025) both investigated LLM’s decision-making capabilities in simplified supply chain games. However, both studies investigate genAI’s ability to make independent decisions, rather than acting as decision support. The studies also take place in solely theoretical contexts, leaving room for exploration in real supply chain settings. Moreover, Novoszel and Claus (2024) developed a framework with the ob-

jective of achieving individualized and customized qualitative and actionable insights by using LLM-based AI in a SCM context. The paper, however, does not include an application of the framework to any real-life supply chain scenario. Therefore, it mainly serves as an indication that the field is relevant yet fairly unexplored.

The relevance of this gap is further supported by Aaijaz et al. (2025), who find that although many organizations recognize the potential of AI for decision-making, they face difficulties in operationalizing it to achieve sustained efficiency. In contrast to traditional AI applications, genAI offers a more interactive and user-oriented approach, with capabilities such as explanation generation and contextual reasoning. These features make genAI particularly relevant for supporting decision-makers, as it can transform complex datasets into interpretable insights.

From an academic perspective, research on digital technologies in SCM is rapidly evolving, with a growing body of literature focusing on AI applications. As noted by Bahroun et al. (2025) the field has seen significant development in recent years. However, there remains a need for empirical case studies that examine how AI creates value in specific organizational contexts. This thesis addresses this gap by investigating the application of genAI in a real-world supply chain setting, thereby contributing both to academic understanding and to practical implementation.

4 Mapping of PipeChain SCM’s Inbound Process

This chapter presents and structures the findings from interviews with PipeChain SCM employees and a customer representative, providing an overview of PC SCM’s inbound process. It also describes the data available in the system that can be interpreted by generative artificial intelligence, followed by possible use-cases. Finally a discussion of what KPIs and deviations should be prioritized when monitoring and analyzing is presented.

4.1 Purpose of Interviews

PC SCM offers a number of applications to their customers, with PC being one of the major ones. The inbound process is one of the system’s major modules, and the process itself varies in configuration depending on customers’ operations. Therefore there is no available standardized definition of how the inbound process is set up, or how customers integrate their own process with the system. Data structure and KPIs monitoring in the system also vary based on customer integration. The first issue to deal with in answering the research questions is therefore to define the case, which is done by mapping the inbound process through interviews with system experts and users.

The goal of the interviews is to develop an understanding of the inbound process, which then, in combination with the literature review, leads to a prompt for a genAI model that prioritizes the same things as system users do in reality. To address the research questions, the process mapping will have to go beyond simply finding how data is structured into KPIs, it will also have to uncover how the process adds value to PC SCM’s customers, and what tasks system users perform in the system.

4.2 Interviewees

Interviews were conducted with eight individuals, seven internal and one external, possessing knowledge and experience of the system in general and the inbound process in particular. All interviewed PC SCM employees hold consulting roles, in which they collaborate closely with customers to ensure that value is created through the use of the system. Table A.1, displaying the interviewees, their official role, and their experience working with the system is found in Appendix A, and the questions used are stated in Appendix B. At first, the seven internal consultants at PC SCM were interviewed and the responses were substantially aligned in a majority of aspects. To validate the findings through triangulation, an external interview was conducted with a PC SCM customer. The insights from this interview with were much alike those from the internal interviews, thus the conclusion was drawn that further interviews were unnecessary.

The interviews were conducted in a semi-structured manner, as described in Chapter 2, Section 2.4.2. During the interviews, emphasis was put on three main areas:

- Understanding the inbound process and how it creates value for PC SCM's customers.
- How the system is used by PC SCM's customers, including typical user roles and how frequently users access the system.
- What data and KPIs are available in the system, and which of these are considered most important to monitor.

Based on the insights obtained from the interviews, a general mapping of the inbound process in the PC system was developed, which is presented in the following section.

4.3 Case-Specific Process

The objective of this section is to provide a general overview of the inbound process, thereby establishing an understanding of the relevance and context of this thesis which is crucial for the development and evaluation of the prompt later on. Rather than offering a detailed analysis of the process, the section aims to present a high-level and accessible description.

PC SCM's inbound process provides a platform for collaboration and visibility between supply chain actors and their suppliers. Both parties have access to the system, enabling shared data and common KPI definitions. The system tracks both physical and financial flows, supporting coordination and information sharing between supply chain actors. While the specific configurations of the process may vary depending on customer requirements, the description below provides a general overview of the inbound replenishment process within the system. Some customers may exclude certain steps, while others may add additional ones. Typical system users include purchasing managers, supply planners, and supplier representatives who monitor deliveries and manage deviations.

From an operational perspective, the inbound process begins when PC SCM's customer places an order to replenish stock. The order specifies the demanded delivery date and the required quantity. When the supplier receives the order, it responds by confirming the delivery date and quantity within PC SCM's system. In addition, a second confirmation is sent prior to shipment. Finally, the physical delivery is received by the customer, who records the real arrival time and delivered quantity in the system. Figure 4.1 displays the information flow in the system for the replenishment process. The confirmations, delivery information, and recorded deviations in the process generate the operational data from which KPIs are calculated within the system.

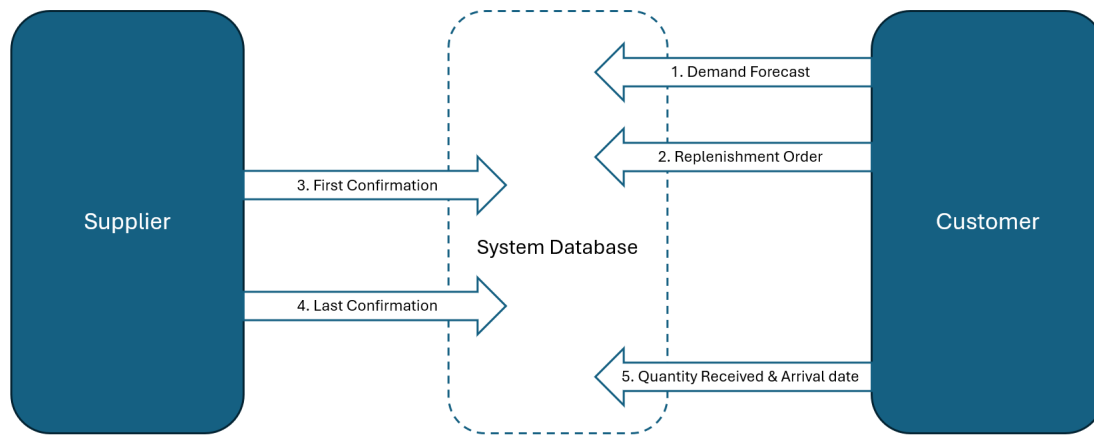


Figure 4.1: *Information flows from replenishment process in the system.*

In addition to the operational replenishment process, the system also supports longer-term planning activities. Customers may share demand forecasts with suppliers over extended time horizons, allowing suppliers to better plan production and deliveries. The forecasting horizon varies depending on the product and supplier, but they typically cover a period of approximately two to four months.

The information generated throughout the replenishment process forms the basis for performance monitoring within the system. Data such as confirmed delivery dates, delivered quantities, and deviations from planned deliveries are recorded and stored over time. This allows customers to evaluate supplier performance using KPIs that can be monitored and compared across suppliers or against predefined targets. Historical data can also be used to analyze performance trends and identify areas of improvement. According to the interviewed customer representative and several consultants, this information is commonly used during quarterly performance meetings between purchasing companies and suppliers.

The system is largely centered around deviation management, as most user actions involve identifying and addressing deviations from expected performance. Users define acceptable thresholds or intervals for deviations, and when these thresholds are exceeded the system generates alerts. Notifications can be sent either via email or as in-system messages, prompting users to investigate and respond to potential supply chain disruptions. As a result, system users primarily interact with the platform when deviations occur, focusing their attention on resolving disruptions when they occur.

As mentioned, the operational data generated within this inbound process forms the basis for the KPI monitoring and analysis that are central to the system's decision-support functionality. For this thesis, genAI will be used to analyze metrics connected to the physical flows between customers and suppliers.

4.4 Available data

The data collected in PC is structured numerical and categorical operational data. Within the traditional replenishment system, two KPIs are primarily displayed: delivery precision (DP) and forecast accuracy (FA). In certain system configurations, service level (SL) is also monitored. These KPIs are calculated using several underlying metrics. DP is determined by tracking three metrics:

- Delivery against demand (D)
- Delivery against first confirmation (FC)
- Delivery against last confirmation (LC)

FA is determined by tracking two metrics:

- Forecast accuracy
- Forecast bias (FB)

In addition to these metrics, the system also collects information such as ordered and delivered quantities, as well as cover time. Overall, a wide range of data is available within the system, at both product and supplier levels.

4.4.1 Delivery precision

DP indicates how often orders are delivered at the confirmed time. For each order line, DP is either 100% when a delivery arrives within the acceptable interval of the confirmed date, or 0% if the delivery arrives outside of this interval. Hence, a supplier's DP is computed as the percentage of shipped rows that has a DP of 100%. Similar logic applies for computing DP for products, sites or product groups. The metric is applied to D, FC, and LC, resulting in three different DP values. LC is the most crucial since that takes place closer to physical deliveries than the others.

4.4.2 Forecast Accuracy and Bias

FA is a metric for how accurate the customer's stated forecast is compared to what they actually order from their supplier. If the value is below what is deemed acceptable, customers are recommended to overlook their forecasting process. FA is computed according to Formula (4.1).

$$FA = 1 - \frac{|Forecast - Actual|}{Forecast} \quad (4.1)$$

FB is another available metric connected to the forecasting process. Similarly to FA, it displays the difference between the forecasted and ordered quantities, with the addition of showing whether forecasts tend to be too high or low. FB is calculated according to Formula (4.2).

$$FB = \frac{Forecast - Actual}{Forecast} \quad (4.2)$$

4.4.3 Service Level

The SL in PC SCM's system can be configured after customer preferences, but the metric is based on four pre-defined intervals of stock levels:

- Blue Interval - Stock level is above acceptable limit
- Green Interval - Stock level is on acceptable level
- Yellow Interval - Low, yet acceptable stock level
- Red Interval - Critical stock level, risk of stock-out

SL is computed as the fraction of time stock-levels spend in non-desired intervals. What intervals are considered non-desired can be configured by the customer, yet the most commonly used set-up is when inventory is in the blue interval, where stock build-up occurs, or red interval, where there is a risk of stock-out. SL as a KPI is mostly used in VMI setups, but is also available in the traditional replenishment process.

4.5 Identifying Points of Interest

As mentioned in Chapter 2, a quasi-static analysis is conducted in the process of identifying the most important topics to be considered when prompting the genAI model. Additionally, frequency is used to determine what issues it should prioritize, since the metrics that were mentioned the most are likely to be used by more customers, thereby increasing the relative importance of the metric. The analysis is performed connected to three main areas:

- Key metrics
- Value Creation
- Usage Areas

4.5.1 Key Metrics

Firstly, KPIs have to be identified and benchmarked in terms of importance to system users. Through the quasi-static approach, four metrics were identified: DP, SL, FA and FB. Figure 4.2 displays these metrics, and in what interviews they were mentioned.

Metric Interview	Delivery Precision	Service Level	Forecast Accuracy	Forecast Bias
#1	☒			
#2	☒		☒	☒
#3	☒		☒	☒
#4	☒	☒	☒	☒
#5		☒	☒	☒
#6	☒		☒	☒
#7	☒	☒		

Figure 4.2: *Frequently mentioned metrics in conducted interviews.*

As seen in the figure, DP was mentioned in all interviews except one, making this the most important metric based on the quasi-static method. Both forecasting metrics FA and FB were mentioned in five interviews, suggesting these are highly important as well. Lastly, SL was mentioned in three interviews, making this the least important, while still important, metric. DP being deemed most crucial can be further supported by two interviewees explicitly mentioning it as the most important, which did not occur for any other metric.

4.5.2 Value Creation

Looking at the process' value creation for customers, Figure 4.3 illustrates the identified key topics from interviewees. These were automation and digitalization, collaboration, and visibility.

Value Creation Interview	Automation / Digitilization	Collaboration	Visibility
#1	☒		☒
#2	☒	☒	
#3	☒	☒	
#4	☒	☒	☒
#5	☒	☒	☒
#6	☒	☒	☒
#7	☒	☒	☒

Figure 4.3: *Frequently mentioned areas of value creation in conducted interviews.*

The figure shows alignment between interviewees regarding the benefits that companies receive from using the system. All interviewees agreed that automation and digitalization was a big part of PC SCM's value proposition. Collaboration between customers and their suppliers was also mentioned frequently, in all interviews except one, and a connection to supply chain visibility and common definitions were emphasized on multiple occasions, having been mentioned in five interviews.

4.5.3 Usage Areas

When investigating how the system and its available data is used, the areas deviation focus, quarterly supplier meetings, and operational order management were identified, and are illustrated in Figure 4.4 below.

Usage Area Interview	Deviation Focus	Quarterly Supplier Meetings	Operational Order Management
#1	☒		☒
#2	☒		☒
#3		☒	☒
#4	☒	☒	☒
#5			☒
#6	☒	☒	☒
#7		☒	☒

Figure 4.4: *Frequently mentioned usage areas in conducted interviews.*

In comparison to the system’s value creation, there is less consensus of how customers are currently using the system. All interviewees mentioned operational order management as an use area, however, only four mentioned deviation focus and quarterly supplier meetings. This is further supported by multiple respondents stating that configurations vary largely depending on customer preferences, meaning that different customers demand different things from the system. Multiple respondents also mentioned that there is room for improvement and opportunities to find new ways to use the available data, especially connected to working proactively and with a more long-term perspective, instead of daily monitoring.

4.6 Takeaways

PC SCM’s inbound process involves traditional supply chain KPIs and concepts. The system offers large amounts of data to its users that can provide both holistic and detailed insights. When letting genAI analyze the data, it will likely identify many issues and trends. Some of which will provide valuable insights while others will prove worthless to system users. Therefore emphasis should be put on providing genAI with a clear description of the data’s context, to ensure that data being interpreted correctly.

Also, the number of issues to be presented by genAI will have to be limited. Several respondents stated the importance of not overwhelming system users with too much information. This in combination with the fact that multiple KPIs are available per supplier, product, and order levels poses a challenge for genAI to prioritize what found issues to display. Consequently, the instructions have to state what KPIs to prioritize higher than others, as well as what level of detail the issues should be

presented in.

In terms of providing genAI with a clear objective, the interviews state digitalization and automation as the primary value creators. If genAI would prove to be proficient in structuring available data, these aspects are enhanced even further. The other found value propositions are collaboration and visibility for supply chain members. Whether or not genAI can improve these aspects depends on the users' purpose of working in the system. Two of the found usage areas are monitoring of daily operations and preparing for quarterly supplier meetings, both of which are deviation-oriented. By including an objective and purpose when instructing genAI on how to analyze the data could increase its ability to provide relevant insights. Hence the prompt will have to provide the context in which the data will have to be presented, including both the environment where the data is collected and the purpose of the analysis.

5 Constructing Foundation for Artificial Intelligence Prompting

In this chapter, a framework for prompting is developed. Firstly, existing frameworks are scanned and discussed with the objective to identify commonalities. Secondly, a theoretical prompting framework is constructed based on the findings from the screening process. The chapter ends with an adjustment of the framework to better align it with the case-specific processes.

5.1 Framework Development Approach

In this thesis, the development and application of a genAI prompting framework is approached through prompt engineering rather than fine tuning. The choice is primarily motivated by the nature of the task and the context of the PC SCM case, which focuses on the interpretation of structured numerical data to identify deviations. Such tasks rely predominantly on logical reasoning, pattern recognition, and rule-based evaluation, rather than deep domain-specific knowledge. As many LLMs already demonstrate strong capabilities in these areas, the additional benefits of fine tuning are assessed to be limited in relation to the required effort, data, and computational resources. Prompt engineering is considered sufficient for the exploratory and iterative nature of this study, where flexibility and rapid adjustment of prompts are prioritized.

Furthermore, the data used throughout the study is fully anonymized, meaning that no sensitive or personally identifiable information is processed. As a result, data security and privacy considerations, while highly relevant in real-world implementations, are not treated as a central constraint in this thesis. This allows the study to focus on methodological development rather than compliance-related challenges, although it should be noted that such considerations would be critical in a real-life environment.

5.2 Existing Frameworks

Before concluding that a new framework needed to be developed, existing frameworks were reviewed with the objective to explore whether an existing framework is applicable to this study. To ensure the relevance of the frameworks, precise search terms were applied. Since this thesis aims to explore the use of AI for numerical data interpretation, it was important that any framework considered could accommodate this aspect. Accordingly, this section of the thesis began with a mapping of currently available frameworks for prompting genAI. As with the literature review, the data base used was Lund University's Finn. The search was conducted across four fields, which are summarized in Table 5.1 below. The first field included *prompt** OR *instruction** and was restricted to titles only. The second field contained *gen** OR *a**, and the third included *framework* OR *model*. Both of these were also limited to titles to ensure that the studies were centered around these main areas. Finally,

to ensure that the framework was at least partially adapted for data handling, the fourth field used *data**, which was not restricted but applied to all fields.

Table 5.1: Search words used for exploring existing frameworks for prompting genAI on Finn.

	Search word	In field
AND	prompt* OR instruction*	title
AND	gen* a*	title
AND	framework OR model	title
AND	data*	all fields

The results were subsequently filtered to include only peer reviewed research in order to enhance validity. After this filtering, only 21 sources remained. To expand the search, the data base Scopus was also consulted. An initial search using the same terms and filters yielded over 1000 articles, indicating the need for a more focused approach. Insights from the Finn search suggested that many articles primarily discuss genAI in the context of image generation, which is outside the scope of this study. To address this, the term image was explicitly excluded from the search to narrow the results. The exact search terms used are illustrated in Table 5.2 below.

Table 5.2: Search words used for exploring existing frameworks for prompting genAI on Scopus.

	Search word	In field
AND	generative artificial intelligence OR genAI	article title, abstract, keywords
AND	prompt OR instruction	article title
AND	framework OR model	article title
AND	data*	article title, abstract, keywords
AND NOT	image	article title, abstract, keywords

The results were further filtered to include only articles within the fields of computer science, engineering, mathematics, or decision sciences, yielding 30 additional articles. Four of these articles had already been identified through Finn and three were inaccessible, resulting in 23 new contributions. Combined with the previous search on Finn, this produced a total of 44 articles and conference papers.

During the review process, it was observed that many sources lacked explicit frameworks or models for prompting. This can largely be attributed to the broad use of the term *model* within AI research. While including *model* as a search term increases the likelihood of identifying studies that propose prompting frameworks, it also increases the risk of retrieving articles that merely reference LLMs or AI models

in general. Consequently, a substantial number of articles were excluded due to lack of relevance. The content of the studies that did present some kind of framework is summarized in Table 5.3 below, where the authors, the name of the framework, and a short summary of the framework is presented.

Table 5.3: Overview of prompt-based frameworks

Reference	Framework Name	Brief Explanation
Kan et al. (2023)	Composable Prompt-based Generative Framework (CPGF)	Decomposes information into fragments and uses modular sub-prompts to construct type-specific prompts for information extraction.
Huang et al. (2026)	Prompt-Guided Generative Language Decoding Model (PG-GVLDL)	Utilizes subject-specific identity embeddings and task-aware prompt descriptors to guide the decoding of neural activity into text descriptions.
Yuan et al. (2025)	UniST (Universal Spatio-Temporal prediction)	Employs a prompt network to adaptively generate spatio-temporal prompts guided by domain knowledge like spatial hierarchy.
Lee and Park (2025)	Structured BAS Prompt Framework	Utilizes a five-component structure; subject, attribute, mood, time/background, and negative prompt, to systematically guide architectural visualizations.
Zhao et al. (2025)	Memory-Driven Prompt Learning (MemPrompt)	Uses generative prompts retrieved from memory and shared prompts to compensate for missing data channels in multi-modal models.
Xu et al. (2024)	EviPrompt	Automatically generates prompts for medical image segmentation based on feature similarity and evidential learning.
Jiang et al. (2025)	Stage-wise Prompt Portfolio Framework	Categorizes prompts into tool-based, collaborative, and mentor-like roles that adapt to different stages of the creative design process.
Jou and Hao (2026)	WSPQCF Framework	Uses AI as a cognitive scaffold for reflection and refinement with strict guardrails against outsourcing actual work.

Continued on next page

Table 5.3: Overview of prompt-based frameworks

Reference	Framework Name	Brief Explanation
Kuligin et al. (2025)	Agentic Workflow Framework	Recommends Reason-Act (ReAct) patterns combined with external tool grounding to optimize specialized medical question answering.
Laeq and Saleem (2025)	SAFE-PE (Systematic Assessment Framework for Evaluating Prompt Engineering)	Provides a standardized five-step assessment pipeline using quantitative and qualitative metrics to evaluate prompt performance and reproducibility.

Although the majority of the reviewed studies emphasize structured, qualitative guidance for task-specific prompting, there is considerable variation in the scope of proposed frameworks. To assess the relevance of existing frameworks in relation to the PC SCM case, a screening process is necessary based on the specific requirements of the application context. As presented in previous chapters, PC SCM requires numerical data interpretation and deviation detection within a supply chain setting. Moreover, explainability for root-causes and actions are important, as well as structured output containing not only text. Thus, particular emphasis is placed on frameworks that enable structured reasoning, transparency, and adaptability to quantitative data.

Accordingly, the screening criteria include the capability to handle numerical or structured data, support for stepwise or explainable reasoning, applicability beyond highly domain-specific contexts, and ability to guide output structure in a way that supports decision-making. Frameworks that did not meet these criteria were considered less relevant for the purposes of this study.

When applying these screening criteria, several limitations of the identified frameworks became evident. A number of frameworks, such as those focused on medical or legal applications, are highly domain-specific and therefore lack transferability to the PC context. Others primarily focus on text-based analysis such as interpretation, classification, and information extraction, or multimodal tasks, with limited emphasis on numerical data interpretation, making them less suitable for numerical deviation detection.

Furthermore, while some frameworks incorporate structured prompting or iterative refinement, they often lack explicit mechanisms for guiding output in a way that supports actionable insights, such as root-cause analysis and recommendation generation. As a result, these frameworks do not fully align with the requirements of the PC case. Consequently, although individual components of existing frameworks are valuable, none provide a comprehensive solution for the targeted application, thereby motivating the development of a new, tailored prompting framework. As shown in Table C.1 in Appendix C, the majority of frameworks score low in terms

of combined numerical handling, explainability, and structured output, further reinforcing the need for a tailored framework.

The screening process thus demonstrates that, while existing frameworks offer useful design principles regarding the construction of prompts for genAI, they do not sufficiently address the combined requirements of numerical data interpretation, deviation detection, and actionable output generation in the PC context. This gap highlights the need for a tailored framework that integrates these aspects into a coherent prompting strategy. Building on these insights, the following section presents the construction of such a framework, specifically designed to support the analytical and decision-making needs of PC SCM.

5.3 Construction of Theoretical Framework

Building on the previous section, the frameworks and guides presented in current research are not directly applicable to this thesis, motivating the construction of a new framework designed for data interpretation and suitable for the PC landscape. This will be done through an iterative process in which the prompt will be evaluated and adjusted. The goal of this framework is to provide the user with a stepwise guide in how to use genAI for interpretation of data and reduce the need for manual workload.

The reviewed literature demonstrates significant variation in how prompting methodologies are designed and applied. Some approaches combine fine-tuning with prompt engineering, while others rely exclusively on prompting techniques. Within the latter category, no single dominant approach emerges, instead prompt complexity ranges from simple zero-shot formulations to more advanced multi-step strategies. This variation highlights that the choice of prompting technique is highly dependent on the specific task and its requirements. In the context of the PC SCM case, which involves structured numerical data interpretation and deviation detection, both accuracy and explainability are critical. Consequently, simpler prompting approaches such as zero-shot or few-shot prompting are considered insufficient, as they may lack consistency and fail to provide transparent reasoning.

A recurring trend in the literature is the use of iterative optimization, where users and LLMs collaboratively refine outputs over multiple steps. This is particularly relevant for the PC context, as it aligns with the need to continuously improve analytical outputs and adapt prompts based on observed performance. These insights suggest that a structured and iterative prompting approach is most appropriate for the task at hand, thereby motivating the design choices underlying the proposed framework.

The findings of Yu et al. (2025) indicate that certain LLMs struggle to perform complex reasoning when relying on simple prompting techniques, suggesting that more advanced approaches, such as CoT prompting, are more effective in such contexts. This is further supported by Vilakati (2025), who evaluated zero-shot, instruction-based, CoT, and hybrid prompting strategies for statistical reasoning. Their results demonstrate that, while zero-shot prompting is easy to apply and sufficient for ba-

sic tasks, it is inadequate for more complex numerical problems requiring advanced analytical reasoning. In addition, zero-shot and few-shot prompting are, in general, less effective for complex analytical tasks as these approaches rely heavily on examples and may introduce variability in the model’s responses. In the context of numerical data interpretation and deviation detection, such variability is undesirable, as consistency and precision are critical. Furthermore, the use of examples does not necessarily enhance the model’s analytical understanding in this setting, which further supports the decision to exclude these techniques in favour of more structured prompting approaches.

Given that the objective is to detect deviations, accuracy and explainability are of significant importance in the construction of the prompting model. It is not sufficient to merely identify the deviations, it is equally important to understand the reasoning behind them. This highlights the suitability of CoT prompting, as it encourages the model to articulate reasoning steps, thereby enhancing transparency and interpretability. In addition, instruction based prompting is considered appropriate for this task, as it provides explicit guidance on how the model should approach the problem, which can improve both accuracy and consistency. Together, these considerations suggest a hybrid prompting approach that combines CoT and instruction based techniques to ensure a high level of both transparency and accuracy. In particular, transparency is reinforced by explicitly specifying the steps the model should follow to arrive at its conclusion.

Furthermore, the study of Vilakati (2025) suggests that a hybrid prompting model is the most effective for handling complex data and tasks, thereby reinforcing the approach adopted in this thesis for framework development. Although many LLMs exhibit strong reasoning capabilities and hold significant potential for data analysis, their effectiveness remains highly dependent on the design and quality of the prompts (Vilakati, 2025). Here, Vilakati (2025) emphasizes the importance of a clearly defined objective to ensure the LLM accurately understands the task. This aligns with a broader consensus in the prompt engineering literature, where clarity and contextual grounding are consistently identified as critical factors for effective prompting. This was also highlighted in Chapter 3, Section 3.2.1. Based on these insights, a well-constructed prompt should clearly articulate its purpose, provide sufficient contextual information, and include explicit instructions on how the task should be performed, along with a precise specification of the expected output. These key components are summarized in Table 5.4 below, presenting a prompting framework based on existing literature.

Table 5.4: The theoretical prompting framework with proposed key components of the prompt

Component	Description
Purpose	Clarifies the intent of the prompt so that the objective is explicit.
Context	Provides relevant background information to ensure the LLM considers relevant factors.
Output specification	Specifies what the LLM should do and how the output should be presented.

5.4 Applying the Theoretical Framework

When applying the framework presented in the previous section to the inbound process of PC SCM, described in Section 4.3, the purpose and context sections of the prompt are intuitive to construct, while specifying output can differ depending on the role of the system user, and the systems set-up. Throughout the prompting tests, a general system integration will be used, where KPIs should be prioritized according to the analysis in Chapter 4. Similarly, the prompt should define what the output will be used for, thereby ensuring relevance of generated insights. The following sections display how the developed theoretical framework is applied to PC SCM's inbound process and further developed to ensure applicability.

Purpose

The purpose part of the prompt refers to one initial sentence explaining why genAI is performing its handed task. In the case of PC SCM, the purpose is to interpret, structure, and present data, thus this is what the purpose section of the prompt should establish. An example, adapted to the PC landscape, is provided below.

Find deviations in the attached data file, analyze and present them both in text and visually.

Context

The next step is to provide context to the model by describing the environment from which attached input data is generated. Describing the process' environment in detail encourages the model to interpret data correctly, as well as generating relevant explanations of deviations. Also, the system user's role is described, thereby eliminating the risk of suggested actions being irrelevant to the user. Finally, available KPIs is defined and explained. The relative importance of metrics is also provided to direct genAI in its analysis, and to state what deviations are more crucial than others. An example, adapted to the PC landscape, is provided below.

The data is collected in a supply chain context, where companies and their suppliers share a common database. Customers place an order with a demanded (D) delivery date. Suppliers confirm the delivery date two

times, first confirmation (FC) and last confirmation (LC). These are compared to D, and delivery precision is computed as a KPI showing how often deliveries arrive at requested dates. DP is measured against D, FC and LC, where LC is the most crucial and D the least. Service level is also available at supplier and product levels. Service level is defined as the fraction of time stock level spends below a crucial threshold. A lacking service level might be a consequence of poor delivery precision. Therefore delivery precision should be prioritized higher when presenting deviations. Deviations are presented to a user who seeks to monitor the purchasing company's inbound order process, and identify which suppliers and products are creating issues.

Output Specification

The output specification part of the prompt should include clear and explicit instructions of how to perform the task followed by how to present it. As established in Section 5.2, a hybrid approach combining CoT prompting with instruction-based prompting is considered the most favorable prompting approach. Thus, this specification should step wise guide the genAI through its task, encouraging it to break down the problem into smaller sub-tasks and accounting for each task separately.

The first sub-task is to go through the data, looking for deviations in available KPIs. Next, the deviations are to be compared and prioritized. As presented in Section 4.3, emphasis is placed on presenting the most grave deviations, rather than all deviations, as the latter risks overwhelming the user. Therefore, only a selected number should be presented. When deviations have been identified and prioritized, they should be presented to the user in an understandable manner, in both text explanations and graphs illustrating the KPI over a specific time period. The text explanation should include a root-cause analysis along with a suggested course of action. An example, adapted to the PC landscape, is provided below.

Go through each row in the three DP columns and mark each row that does not consistently have 100% as a deviation. Provide text-explanations of the found deviations and illustrate each deviation in a graph. If there are more than three deviations, only present the three that are most severe. For the presented deviations, do a root-cause analysis and based on this give suggested recommendations for the user.

5.5 Restructuring the Framework

A pilot test of the theoretical framework revealed that a model solely based on existing literature is insufficient when applied to PC SCM's inbound process. The primary issue identified was related to the structure of the output, where the genAI produced redundant information and provided insufficiently structured analysis. Furthermore, the prioritization logic of the genAI lacked transparency, indicating that the framework did not effectively facilitate the CoT approach. Overall, these findings suggest that a refined framework is required, taking these limitations into

account, and working in an iterative process in which the genAI is prompted and the output evaluated, leading to changes in the initial prompt.

To better support the CoT approach, the prompt design was revised from a continuous, fluent text format to a stepwise instruction structure, which is expected to improve the model’s comprehension and reasoning process. In addition, issues related to output specification were addressed by separating this component into two distinct sections. This adjustment was based on the assumption that clearer structural separation would enhance the model’s ability to follow the intended reasoning process. Consequently, a dedicated task explanation component was introduced to the prompt framework. Rather than combining task description and output specification within a single section, these elements are now treated independently, ensuring that sufficient and explicit guidance is provided to generate the desired output. Examples of generated output by both the original and adjusted frameworks are shown in Appendix D and Appendix E respectively.

Task Breakdown

The task breakdown should include a step-wise guide of the task at hand. By explicitly breaking down the task and stating each sub-task step-wise, it becomes easier to ensure that the genAI is doing what is intended, not focusing on undesired tasks or calculations. In contrast to the theoretical prompting model, consisting only of the output specification, this task breakdown should include an even more detailed description of the task, and no output specifications. An example, adapted to the PC landscape, is provided below.

1. *Go through each row in the three DP columns.*
2. *Mark each row that does not consistently have 100% as a deviation.*
3. *Note the suppliers that are experiencing deviations*
4. *Note the poor-performing suppliers and note what products are experiencing deviations.*
5. *Do a root cause analysis for the deviations.*
6. *Based on the findings and analysis, give recommended actions for the user.*

However, solely excluding the task description from the output specification is not enough, as the output itself lacked in quality as well. Therefore this part of the prompt is adjusted as well, to be even more precise with the objective of eliminating the redundant information currently presented, continuing with the step-wise instructions approach used for the task breakdown. An example, adapted to the PC landscape, is provided below.

The findings should be presented in a powerpoint:

Cosmetics:

Black and green colors as theme of the presentation

All graphs should have clear axis headers

Structure:

Slide 1: Front page titled “Deviation Report” presenting:

- 1. Number of rows analyzed*
- 2. Number of deviation rows*
- 3. Overall DP(LC) percentage*
- 4. Number of poor performing suppliers*

Slide 2: Summary of supplier performance visualizing poor performing suppliers and summarizing the analysis containing:

Overall KPI analysis on supplier level

Bar chart illustrating the poor performing suppliers.

X-axis: supplier

Y-axis: number of deviations

Highlight the most critical suppliers (maximum 5) listing:

- 1. DP(LC)*
- 2. Criticality state and short description of why*

Supplier slides (Number varies based on number of deviations. One slide per supplier): Presentation of each of the deviations, one slide per deviation. The format is:

Header Format: “Supplier name - Issue summary”

Present why this supplier is identified as problematic through KPI visualization of the available time period in data.

If there is a pattern of certain products causing a large part of the supplier’s issues, present those products.

Present and motivate possible root cause(s) of the issues.

Suggest possible actions to manage issues

Final slide: Summarize recommended actions for each of the suppliers

The actions should be divided into immediate, short-term and long-term, displayed as three different columns.

In each column actions for each poor performing supplier should be stated as

“Supplier X: recommended action in one sentence”

Input Data Explanation

Additionally, it was suspected to be necessary to provide further explanations of the input data, as it may lead to less redundant information in the output if already presented in the prompt. However, as mentioned in Section 3.2.1 it is necessary to ensure only relevant information is presented in the prompt to ensure the genAI do not get confused about the purpose and the task at hand. This part of the prompt also ensures that the data file do not need any adjustments before prompting. An example, adapted to the PC landscape, is provided below.

The data shows placed orders over time, with each row representing one order row. Data can be analyzed on four different levels; supplier, product group, product and order. Product groups and products are sub-groups to suppliers. Time is irrelevant in the column for dmd shipment time.

5.6 Final Framework

These presented additions to the prompt result in a refined framework structure consisting of five parts rather than three; purpose, context, input data explanation, task breakdown, and output specification. These five parts are illustrated in Table 5.5.

Table 5.5: The adjusted prompting framework with proposed key components of the prompt

Component	Description
Purpose	Clarifies the intent of the prompt so that the objective is explicit.
Context	Provides relevant background information to ensure the LLM considers relevant factors.
Input Data Explanation	Describes the input data and what it is stating, minimizing the amount of redundant information in the output
Task Breakdown	Divides the task into clear and numbered sub-tasks ensuring transparency and control of the LLM
Output specification	Specifies what the LLM should do and how the output should be presented.

The framework developed in this chapter, and illustrated in the table above, establishes a structured approach to prompt design, grounded in both existing research and the specific requirements of numerical data interpretation. However, as highlighted throughout the literature, the effectiveness of prompting strategies is highly context-dependent and cannot be fully determined from a conceptual perspective alone.

To assess its practical relevance, the framework must therefore be applied and evaluated in a real-world setting. Such evaluation not only enables validation of its performance, but also provides insights into potential limitations and areas for refinement.

The next chapter addresses this by implementing the framework within the PC SCM context, where it is tested, iteratively adjusted, and evaluated using both quantitative and qualitative measures.

6 Evaluation of Framework

In this chapter performance evaluation of the revised frameworks is defined and performed. The framework used for prompt evaluation is initially presented, followed by descriptions of the six aspects that are evaluated through analysis of produced output. Finally, test results are presented.

6.1 Testing

Thorough testing is required to explore the quality of prompts constructed through the adjusted framework. As presented in Section 3.2, Ramamoorthy (2025) emphasizes the importance of both qualitative and quantitative evaluation. Thus, to evaluate the output holistically, detailed tests are performed, taking several aspects into consideration. As presented in Section 5.2, the SAFE-PE (Systematic Assessment Framework for Evaluating Prompt Engineering in Generative AI) framework developed by Laeeq and Saleem (2025) presents an evaluation approach for prompts. This will be used for evaluating the prompts constructed in this thesis. The framework has a holistic approach in evaluating prompts, aligning well with the creation of insights suitable for the thesis' research questions as well as the importance of including both qualitative and quantitative aspects. The SAFE-PE methodology is presented in Figure 6.1 below.

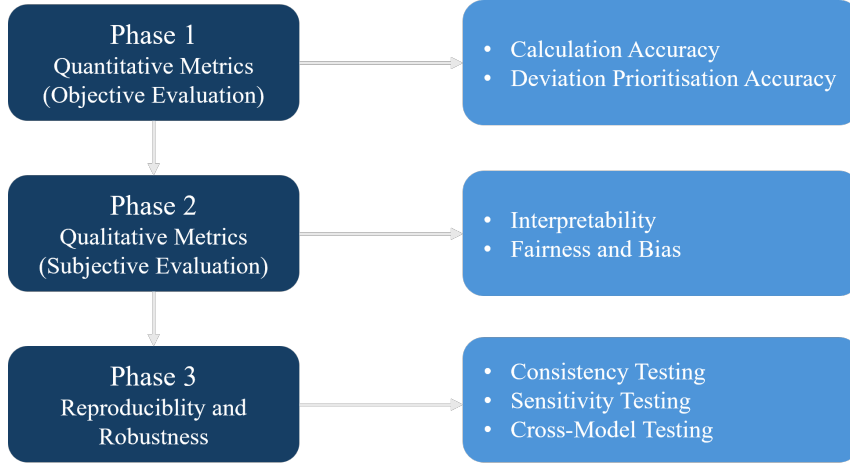


Figure 6.1: *An adaptation of the SAFE-PE methodology presented by (Laeq and Saleem, 2025), adapted to this thesis*

In accordance with Laeeq and Saleem (2025), the prompt evaluation is structured into three phases. The first phase, objective evaluation, focuses on quantitative metrics. By comparing the generated outputs with established benchmarks or blueprints, task accuracy is assessed. In addition, this phase evaluates aspects such as output diversity and efficiency. The second phase, subjective evaluation, involves human judgment and addresses dimensions that are difficult to capture through

quantitative measures alone. Particular emphasis is placed on interpretability, clarity, fairness, bias, and domain relevance. Although Laeeq and Saleem (2025) also highlights the importance of ethical alignment in this phase, ethical considerations are excluded from this thesis, as outlined in Chapter 1, Section 1.5. Therefore, this aspect of the framework will be ignored. The third and final phase concerns reproducibility and robustness, where the stability and generalizability of the findings are examined. This is achieved through sensitivity testing and cross-model testing.

Applying the SAFE-PE framework to evaluate output quality from genAI produces a detailed, yet holistic, overview of the output’s strengths and weaknesses. However, performing evaluation on one single process configuration entails a high dependency on that set-up, as well as KPIs in that particular process being interpreted correctly. Consequently, the SAFE-PE framework will be applied to a second process configuration, where a different set of metrics and a new aim of the analysis are included in the prompt. Identical tests are performed for both process configuration, thereby increasing the robustness of the results. Consequently, most tests were performed on two different data sets describing two different environments: environment 1 (E1) and environment 2 (E2). E1 consists of DP and SL data, however, the analysis is focused on the DP data. This environment can be described with the following case-description:

A system user wants to review their supplier’s performance, with a focus on identifying which ones are problematic, as well as what is causing any issues. The analyzed data includes purchase orders, quantities, and requested delivery dates over a 12-week period. The customer does not have a VMI flow, and the setup is made without any forecasting process. The available KPIs are delivery precision and service level.

E2 consists of service level and forecast data for a production plant with two sites, and the analysis includes both of the KPIs. This environment can be described with the following case-description:

A system user who wants to review their production planning in order to identify problematic products. The data shows production orders over one year for two production facilities. The available data includes service level for all products. Forecast values (forecast accuracy and bias) are also available for some products, but this setup has not been implemented across all products.

Based on the SAFE-PE framework, a total of six tests with the objective of evaluating the prompt are constructed and further explained in the following sections. An overview of the tests and their characteristics are provided in Figure 6.2 below.

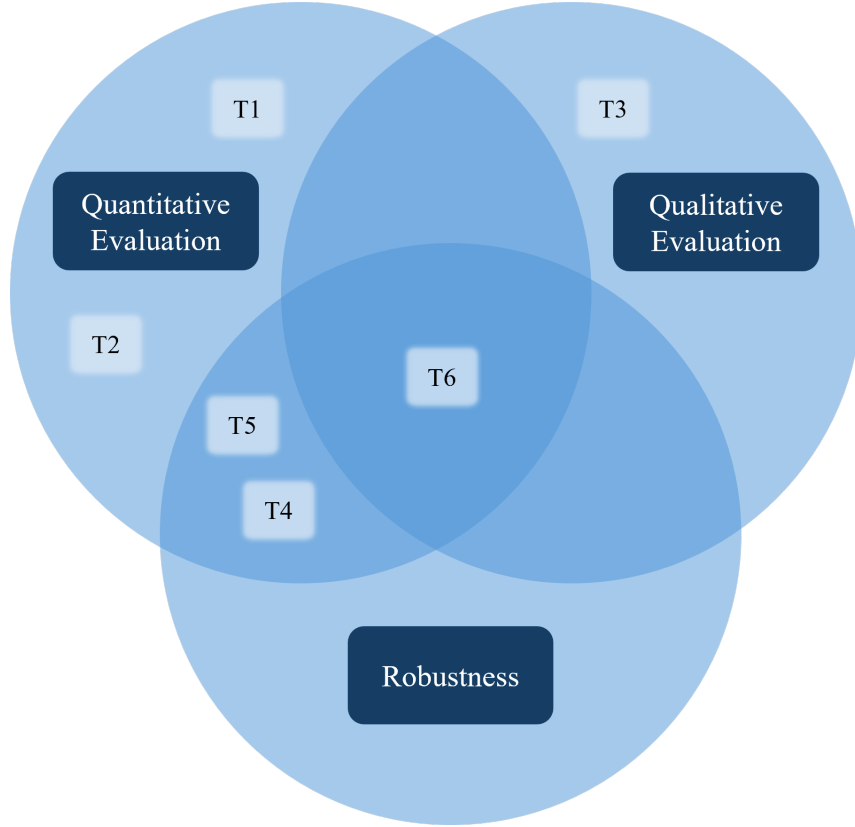


Figure 6.2: Overview of evaluation of test-results based on the SAFE-PE framework.

The main objective of the testing is to validate the prompting framework, and this way address the stated research questions for this thesis. Table 6.1 presents the six tests and the research questions they are connected to.

Table 6.1: Conducted tests using the SAFE-PE framework and the research questions connected to each test.

Test	Evaluation Aspect	Research Questions
T1	Calculation Accuracy	RQ1, RQ2
T2	Deviation Prioritization Accuracy	RQ1, RQ2
T3	Expert Assessment	RQ3
T4	Consistency of Deviation Reliability	RQ1, RQ2, RQ3
T5	Sensitivity Testing	RQ1, RQ2, RQ3
T6	Cross-Model Testing	RQ1, RQ2, RQ3

Every test and iteration was conducted by creating a new chat in the genAI model, reducing the risk of earlier tests affecting the output. For each test, every genAI model was prompted with identical prompts and Excel files containing the data to be analyzed was attached in the chat.

6.1.1 Phase 1 - Quantitative Metrics

As previously outlined, this phase focuses on evaluating the accuracy of the generated output. Within the objective output evaluation, multiple tests are conducted. Specifically two quantitative assessments are performed: calculation accuracy (T1) and deviation prioritization accuracy (T2). Both tests provide empirical evidence that contributes to addressing research questions 1 and 2.

T1: Calculation Accuracy

To determine the calculation accuracy (CA), numerical values presented by genAI are checked by manually calculating the same numbers. The number of correctly presented values by genAI are then divided by the total number of numerical values presented, as described in Formula 6.1. The fraction between these two indicates the accuracy of presented KPIs.

$$CA = \frac{\text{Number of Correctly Computed Values}}{\text{Total Number of Presented Values}} \quad (6.1)$$

A low score suggests a high influence of hallucinations or other reliability issues, whereas a high score suggests accurate results and high model reliability.

T2: Deviation Prioritization Accuracy

To test the deviation prioritization accuracy (DPA) the number of correctly presented deviations is manually checked as in T1. This is then divided by the total number of presented deviations, as described in Formula 6.2.

$$DPA = \frac{\text{Number of Correctly Presented Deviations}}{\text{Number of Presented Deviations}} \quad (6.2)$$

For T2 the prompt is adjusted to include more precise instructions for how to find deviations in the provided dataset. The given search-logic to be included in the task-breakdown section of the prompt, for E1 and E2 respectively, is stated below.

Environment 1: Deviations should be prioritized based on how many rows in data the supplier or product are performing poorly in. The suppliers with the most deviating orderrows are the ones to present.

Environment 2: Deviations should be prioritized based on the number of weeks the product's service level is below 95%. Products with the most weeks with poor service are the ones to present.

A high score indicates reliable outputs as it suggests that the AI is able to correctly prioritize deviations, whilst a low score on the other hand indicates that the genAI lacks in accuracy regarding finding deviations in the data.

6.1.2 Phase 2 - Qualitative Metrics

As presented above, in contrast to Phase 1, Phase 2 involves human judgement to assess the interpretability, fairness, and relevance of the output. One qualitative assessment is performed: expert assessment (T3). This test provides empirical evidence that contributes to addressing research questions 3.

T3: Expert Assessment

In the qualitative evaluation of the output, emphasis is placed on systematically assessing key aspects to determine how effectively the genAI interprets the case process. The evaluation focuses on four metrics:

- Relevance of root causes
- Relevance of suggested actions
- Process interpretation
- Value creation

These metrics are assessed and evaluated through expert judgment inspired by the methodology used by Yeen et al. (2024) and presented in Section 3.2. The same interviewees involved in the data collection presented in Chapter 4 are consulted for this evaluation, as they possess in-depth knowledge of the process and are therefore well positioned to evaluate the relevance and quality of the output. Each expert evaluates the four metrics using a five-point scale. Table 6.2 summarizes the qualitative evaluation framework.

Table 6.2: Expert assessment areas for qualitative prompt evaluation.

Area of Evaluation	Question
1 Process Interpretation	Are the comments in the report appropriate with regard to the inbound process?
2 Root-Cause Suggestions	Are the stated root causes motivated, realistic, and relevant?
3 Suggested Actions	Are the suggested actions for handling deviations motivated, realistic, and relevant?
4 Value Creation	Is the report relevant and does it support customer decision-making?

The experts are instructed to evaluate the generated qualitative output with the criteria outlined in Table 6.2 in mind. For the test, one output from T1 in E1 and E2 from each model are handed to the experts for assessment. The experts are encouraged not to assess whether the output accurately reflects a typical process setup, but rather to focus on how effectively the genAI motivates, formulates, and justifies its analysis. Furthermore, the respondents were unaware what model had produced what report to ensure a bias free evaluation. To provide the experts with a context, the following descriptions of the processes configuration was used:

Environment 1

A system user want to review their production planning in order to identify problematic products. The data shows production orders over one year for two production facilities. Available data includes the service level for all products. Forecast values (forecast accuracy and bias) for some products are also available, but this dataset has not been created for all products.

Environment 2

A system user wants to monitor its production scheduling process to identify problematic products. The data contains production orders over a 6-month period regarding two production plants. Service level is available for all products, with some also having a forecasting process integrated, making forecast accuracy and bias available.

A high aggregate score indicates strong relevance and interpretability, suggesting that the genAI overall generates suitable text given the analyzed data's context, whereas a low score reflects the opposite.

6.1.3 Phase 3 - Robustness Metric

Robustness concerns the stability and generalizability of the findings. In evaluating robustness, emphasis is placed on consistency of the generated output and the sensitivity of identified deviations to variations in the prompt. As previously noted, this can be assessed through sensitivity testing and cross-model testing, but also through testing the consistency of the deviations (T4) by iteratively prompting the genAI with the same data and prompt and noting if any differences in output occurs.

In sensitivity analysis (T5), minor modifications are introduced to the prompt to examine whether such changes lead to variations in the output. Cross-model testing (T6), in turn, involves applying the same prompt and input data across multiple genAIs and analyzing differences in their outputs. This phase addresses the requirements for reliability and validity, as outlined in Chapter 2 and provide empirical evidence that contributes to addressing all three research questions.

T4: Consistency of Deviation Reliability

The consistency of deviation reliability (CDR) of the model is calculated by checking the output after each iteration, and comparing it to the outputs from other iterations. To calculate this metric, five iterations are performed and the number of different outputs are divided by the number of iterations. Formula 6.3, describes how the CDR is calculated, in which the number of unique deviation sets is the number of different outputs generated by the genAI, and the number of prompt iterations is the number of times the genAI is prompted with the same prompt.

$$CDR = 1 - \frac{\text{Number of Unique Deviation Sets}}{\text{Number of Prompt Iterations}} \quad (6.3)$$

A score of 1 indicates that the prompt results in reliable outputs as it generates the same output no matter the number of iterations. Lower scores indicate larger variation in what set of deviations genAI choose to present.

T5: Sensitivity Testing

The sensitivity (S) is evaluated by introducing minor variations to the prompt and assessing the extent to which the outputs differ as a result. The modifications do not alter the substantive content of the prompt but instead involve slight changes in formulation and presentation. In total, three such change modifications are applied, resulting in three distinct prompt iterations. These alterations are displayed in Appendix G. The sensitivity metric is defined as in Formula 6.4, in which the number of unique deviation sets are the number of different output variations received from the genAI, and the number of prompt iterations is the number of different prompt modifications used in the test.

$$S = 1 - \frac{\text{Number of Unique Sets}}{\text{Number of Prompt Iterations}} \quad (6.4)$$

A score of 1 indicates a robust prompt that is resilient to minor variations, consistently producing identical deviation sets across iterations. Conversely, lower scores indicate reduced robustness, reflecting greater variability in outputs due to small changes in prompt formulations.

T6: Cross-Model Testing

When conducting cross-model testing, four genAI models are prompted and evaluated. The models are MS Copilot, as PC SCM, like many organizations, use the MS landscape, ChatGPT as this is the most commonly used genAI, and Gemini and Claude as these also tops the lists of most commonly used genAI models (Moore, 2026; Olavsrud, 2025). The exact versions of the models used for testing are presented in Table 6.3 below.

Table 6.3: Versions of the genAI models used for testing and their originators.

Originator	Version
Anthropic	Claude Opus 4.7 (Adaptive Thinking)
OpenAI	ChatGPT 5.5 Plus
Google	Gemini 3.1 Pro
Microsoft	Copilot GPT-5.3 Instant

As seen in Table 6.3, the models used for testing are developed by different companies and are therefore most likely trained in different ways. However, all four companies state that their model is capable of conducting data analysis. A deeper inquiry of the differences of the models are not made in this study.

By conducting cross-model testing, the outputs provided by the different models can be evaluated and compared to each other. This way, it becomes possible to detect whether one or more models are superior to others regarding the context of PC SCM. Additionally, a consistent result across different genAIs will suggest generalizability of the prompt.

6.2 Results

The tests described in the SAFE-PE framework presented by Laeeq and Saleem (2025) were conducted on the four different genAI models previously presented. The results are presented grouped by their main characteristic: quantitative, qualitative, or robustness. This step addresses the fourth step of the DSR framework: demonstration.

Before conducting the tests, a manual check of both environments were made with the objective of identifying the top ten most critical issues. The results for E1 are illustrated in Figure 6.3 below.

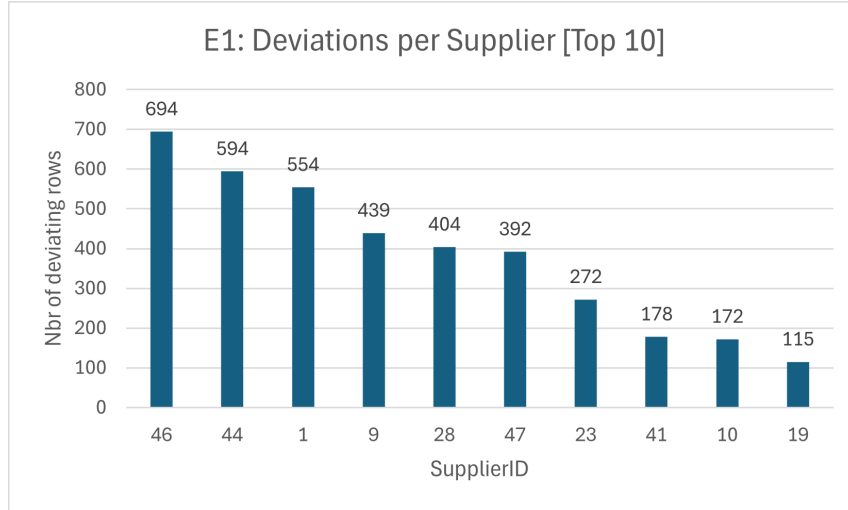


Figure 6.3: *The ten suppliers with the most deviating rows in E1.*

In E1, a deviation is defined as at least one of the three DP columns in a order row containing a zero. Thus, the suppliers are sorted by the most order rows containing an issue connected to at least one of the three DP measurements. As seen in the figure, there is a big gap between the most critical supplier and the second most critical. The next big gap occurs between the third and fourth most critical supplier, and there is another between the sixth and the seventh. The last three suppliers in the top ten are experiencing a fraction of the deviations that the most critical supplier is.

The results for E2 are illustrated in Figure 6.4 below.

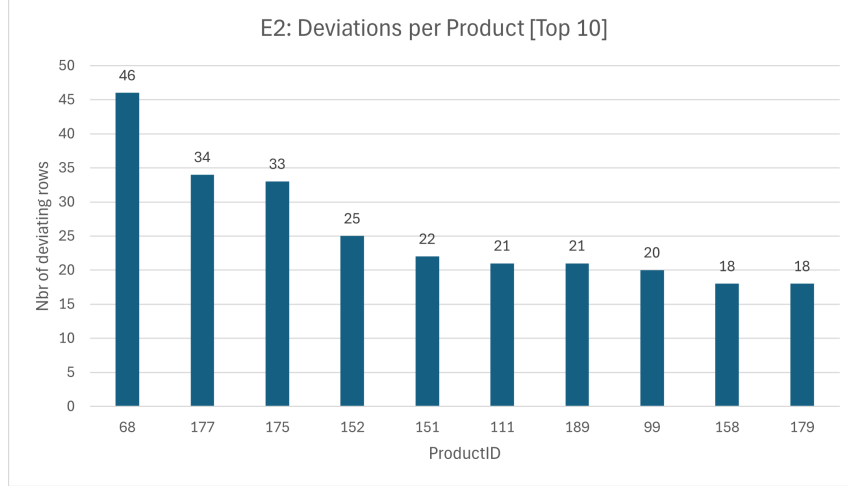


Figure 6.4: *The ten products with the most deviating rows in E2.*

In E2, a deviation is defined as a product experiencing a SL below 100 percent. Thus, the products are sorted by the number of weeks experiencing a service level below 100 percent. As seen in the figure, the most crucial product is clearly distinguished. The following two products are both experiencing similar criticality, whilst the fourth most critical product can be somewhat distinguished from the third and fifth. However, in regards to the next four following products the gap between the number of deviations is very small, the deviations being 22, 21, 21, and 20 respectively. The last two in the top ten both face the same number of deviations.

As the aspect of interest is deviation detection, emphasis is put on accuracy. If a model does not score well in this aspect, it is of no use for the purpose of detecting deviations and supporting decision-making. Therefore, the criteria for each tested model is to achieve at least 50% in CA in at least one test environment for further testing to be conducted on the model.

6.2.1 Phase 1 - Quantitative Metrics

In addition to the results from T1 and T2, the total number of generated values were collected to further get an understanding of the quantitative capabilities of the genAI models. These results have been gathered in Table 6.4 below.

Table 6.4: Summary of the number of generated values and results from T1 and T2 in both test environments for all genAI models.

GenAI Model	Generated	Values	T1 (CA)	T1 (CA)	T2 (DPA)	T2 (DPA)
<i>Test Environment</i>	<i>1</i>	<i>2</i>	<i>1</i>	<i>2</i>	<i>1</i>	<i>2</i>
Claude Pro	122	40	100%	100%	100%	100%
ChatGPT Plus	135	79	100%	94.9%	100%	100%
Gemini Pro	33	14	100%	6.25%	100%	20%
MS Copilot	8	1	25%	0%	—	—

As illustrated in the table, MS Copilot generated the fewest values, with a total of eight. It can also be identified as the least accurate model with respect to CA, scoring only 25 percent. Due to this comparatively poor performance, MS Copilot was excluded from subsequent tests. This decision was based on the premise that results from later tests would be of limited relevance if the model demonstrates insufficient capability in CA, thereby failing to provide reliable data for further analysis.

ChatGPT generates the highest number of values in the output report; however, it does not consistently have 100 percent CA, reaching only 94,9 percent when evaluated in E2. The inaccuracies regard the calculation of FA. Claude produces a similar number of values compared to ChatGPT but is consistent in delivering 100 percent accuracy in both environments. Gemini Pro generates substantially fewer values, approximately 90 fewer than Claude and more than 100 fewer than ChatGPT. Furthermore, in E2 Gemini is failing completely in terms of accuracy, achieving only 6,25 percent CA. Here, most of the inaccuracies are connected to the failure of finding correct deviations. Notably, both Claude and ChatGPT demonstrate 100 percent DPA in T2, whilst Gemini only reaches this score in E1 and drops to 20 percent in E2. This test was not conducted on MS Copilot.

6.2.2 Phase 2 - Qualitative Metrics

As previously mentioned, one report generated from T1 in E1 and E2 was selected for each genAI model. The expert assessments were initially intended to be conducted by the interviewees who contributed insights to Chapter 4. However, due to time constraints, this was not feasible. Consequently, the assessments were instead conducted by the researchers.

Each presentation was evaluated separately and independently by the researchers, and the detailed results are presented in Appendix F. Table ?? below summarizes the average score for each assessed aspect across the evaluated genAI models in E1. As presented in Table 6.2, area of evaluation 1 is the process interpretation, 2 is root-cause suggestions, 3 is suggested actions, and 4 is value creation.

Table 6.5: The average results from the assessments conducted in T3 for E2.

GenAI Model	Process Interpretation	Root-cause Suggestions	Suggested Actions	Customer Value	Average
Claude	5	5	4,5	5	4,875
ChatGPT	4,5	3,5	3,5	3,5	3,75
Gemini	3	1,5	1	1,5	1,75

Table 6.6 below summarizes the average score for each assessed aspect across the evaluated genAI models in E2.

Table 6.6: The average results from the assessments conducted in T3 for E2.

GenAI Model	Process Interpretation	Root-cause Suggestions	Suggested Actions	Customer Value	Average
Claude	5	4,5	3,5	5	4,5
ChatGPT	3,5	4,5	3	4	3,75
Gemini	1,5	1	1	2	1,375

As shown in both tables, Claude achieved the highest overall score among the evaluated models. It received either the highest or an equal score compared to the other models across all assessed aspects, particularly in process interpretation and customer value where it achieved the highest possible score. ChatGPT achieved the second-highest overall score and matched Claude in root-cause suggestions, although it received lower scores in the remaining categories. Gemini demonstrated the lowest performance across all evaluated aspects, with substantially lower average scores than both Claude and ChatGPT. Looking at the two test environments, both Claude and Gemini yield higher average scores in E1, while ChatGPT achieves equal average scores. However, there are no substantial differences between the scores of any model in E1 compared to E2.

6.2.3 Phase 3 - Robustness Metrics

The results from T4 and T5 are presented in Table 6.7 below. When conducting T4, the same prompt as in T1 was used. The modified prompts used in T5 are, as previously mentioned, provided in Appendix G. The prompt was revised three times for each of the two environments, resulting in a total of six different prompts variations.

Table 6.7: Summary of the results from T4 and T5 in both test environments for all genAI models.

GenAI Model	T4 (CDR)	T4 (CDR)	T5 (S)	T5 (S)
<i>Test Environment</i>	<i>1</i>	<i>2</i>	<i>1</i>	<i>2</i>
Claude	1	0,2	1	1
ChatGPT	0,6	0,6	1	0,33
Gemini	1	0	0	0

In terms of CDR-values computed in T4, the values received from testing in E1 range between 0,6–1, while E2 has corresponding values between 0 and 0,6. Both Claude and Gemini perform equally in E1, providing the exact same values in the output for all five iterations resulting in the score of 1. ChatGPT, on the other hand, provides the same output values four out five times in both environments, resulting in the score of 0,6. In E2, Gemini completely fails with a score of 0, whilst Claude also drastically deteriorates, providing the same output only two out of the

five iterations resulting in the score of 0,2. Figure 6.5 below presents the number of presented suppliers for each of the five iterations in E1, and for each of the three genAI models tested.

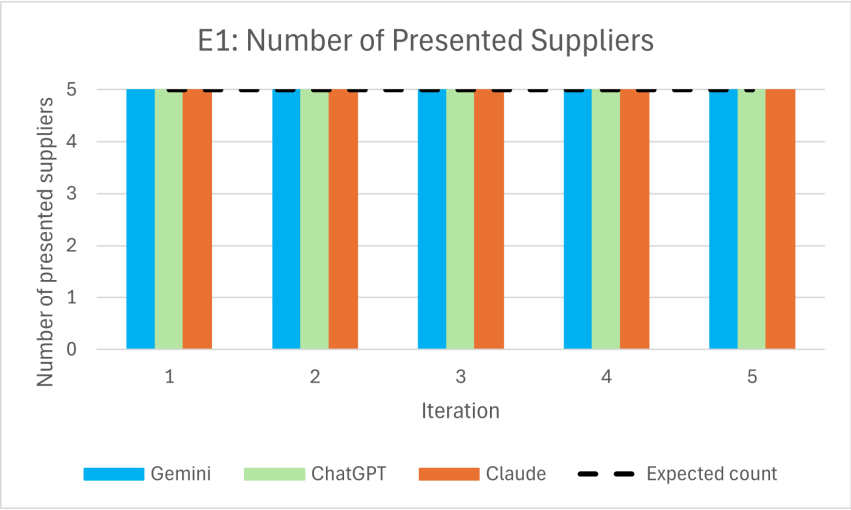


Figure 6.5: *The number of presented suppliers by Gemini, ChatGPT and Claude, for each of the five iterations in T4, E1 along with the expected number.*

As can be seen in Figure 6.5, all three models consistently provided the desired number of suppliers each iteration. However, when it comes to what suppliers were presented the models differ as can be seen in Figure 6.6 below.

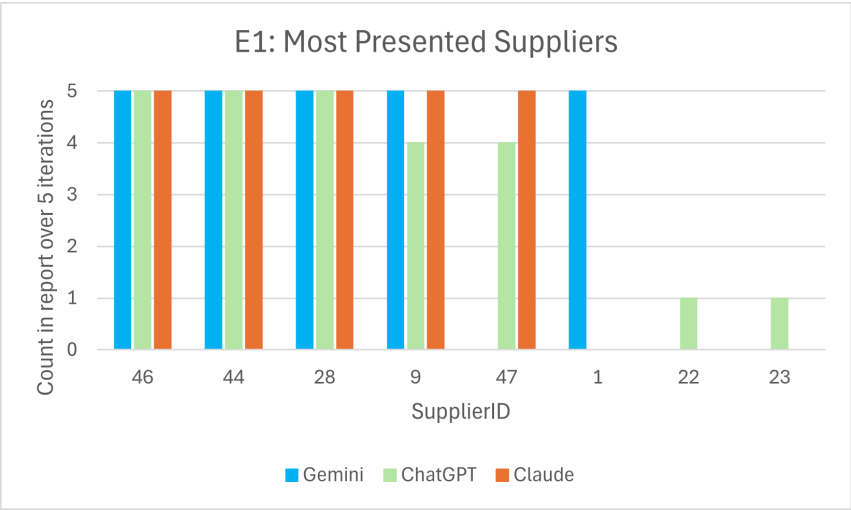


Figure 6.6: *All presented suppliers and their frequency of being deemed one of the five most problematic suppliers by Gemini, ChatGPT and Claude in T4, E1.*

Gemini and Claude are consistent in what suppliers they present as the most critical over all iterations, however, they do not present the same five suppliers. They both highlight suppliers 46, 44, 28, and 9, but then Claude distinguished supplier 47 as critical whilst Gemini emphasizes supplier 1. In four out of five iterations ChatGPT

agrees with Claude, but in one iteration suppliers 9 and 47 are switched out for suppliers 22 and 23.

As previously mentioned, analysis in E2 concerns products instead of suppliers. Figure 6.7 below illustrates the number of presented products for each iteration and genAI model.

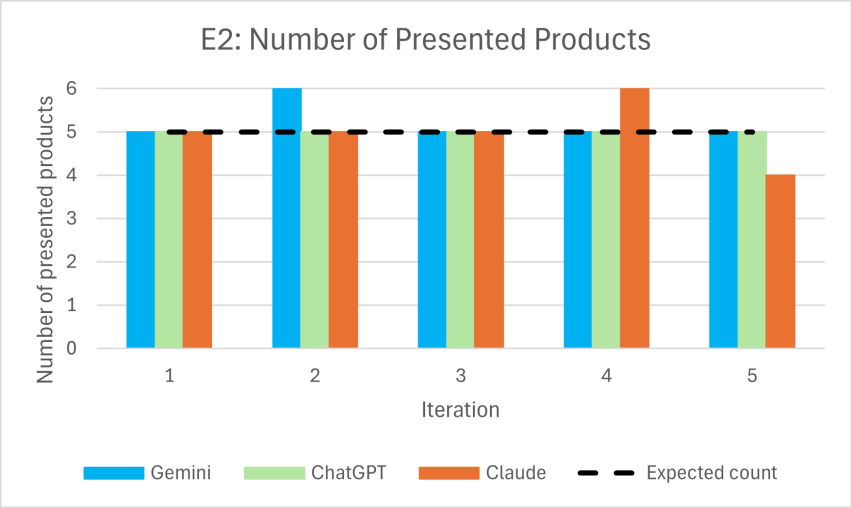


Figure 6.7: *The number of presented products by Gemini, ChatGPT and Claude, for each of the five iterations in T₄, E2 along with the expected number.*

As can be seen in Figure 6.7, ChatGPT is the only model that consistently presents the desired number of products. Gemini presents one additional product in iteration 2, whilst Claude does this in iteration 4 and then presents on less product than desired in iteration 5. Regarding what products are presented, this too differs between the different models, as can be seen in Figure 6.8 below.

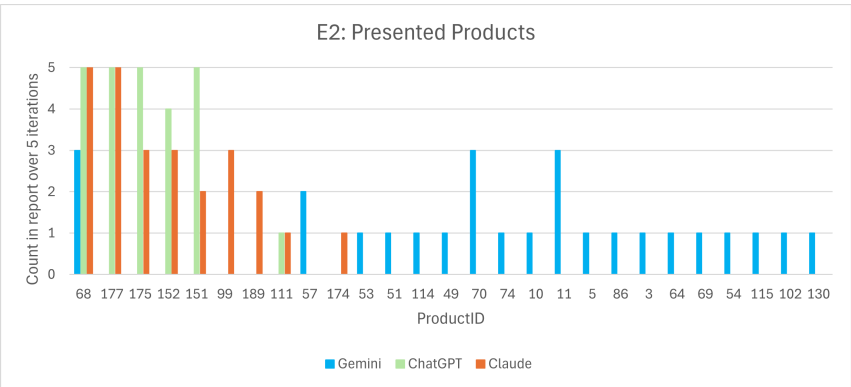


Figure 6.8: *The presented products by Gemini, ChatGPT and Claude during T₄, E2 and their frequency of being deemed one of the five most problematic issues.*

Looking at Figure 6.8, it can be distinguished that ChatGPT is the model that has the least spread regarding the products presented during the five iterations. Product

68, 177, 175, and 151 are consistently presented as critical in all five iterations. In four out of five iterations product 152 is emphasized as critical, whilst in one iteration this has been switched to product 111. Claude presented products 68 and 177 all five times, products 175, 152, and 99 three times, products 151 and 189 two times, and products 111 and 174 one time. Out of these products, all except product 174 are in the top 10 products with the most deviations. Gemini does not highlight a single product as critical all five times, but instead has product 68, 70, and 11 as the most frequently emphasized products having highlighted these three out of five times. Product 57 is distinguished as critical two times, and 15 other products are highlighted once.

Referring back to Table 6.7, when assessing the sensitivity to prompt modifications Claude provided the same output no matter the prompt, resulting in a score of 1. Gemini on the other hand, fails to provide a single duplicate output, resulting in the score of 0. ChatGPT performs well in E1, with a perfect score of 1, whilst producing a unique answer to prompt modification two in E2, resulting in a score of 0,33. Figure 6.9 below presents the number of presented suppliers for each of the three iterations in E1, and for each of the three genAI models tested.

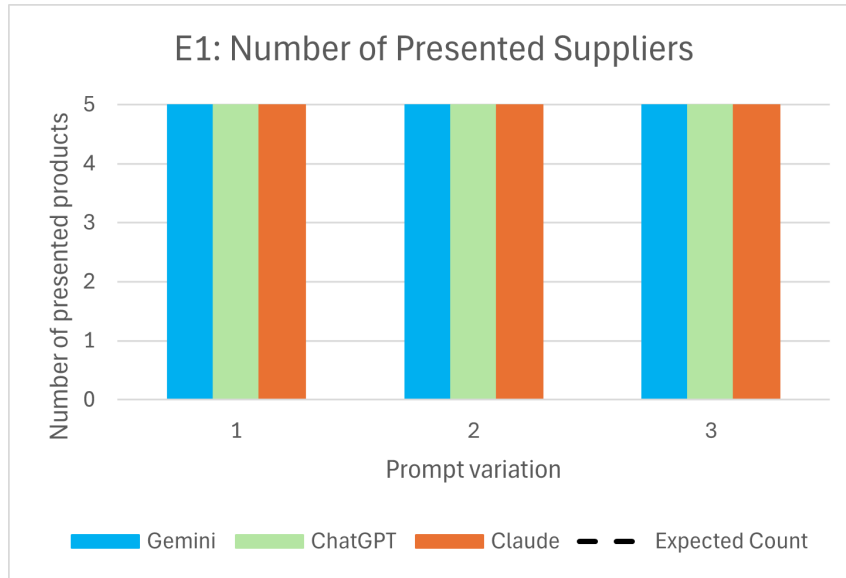


Figure 6.9: *The number of presented suppliers each iteration by Gemini, ChatGPT and Claude for T5, E1, along with the expected number.*

As can be seen in the graph, all three models consistently present the expected number of suppliers each iteration. However, regarding what suppliers are emphasized the models differ in a similar way as in T4. This is illustrated in Figure 6.10 below.

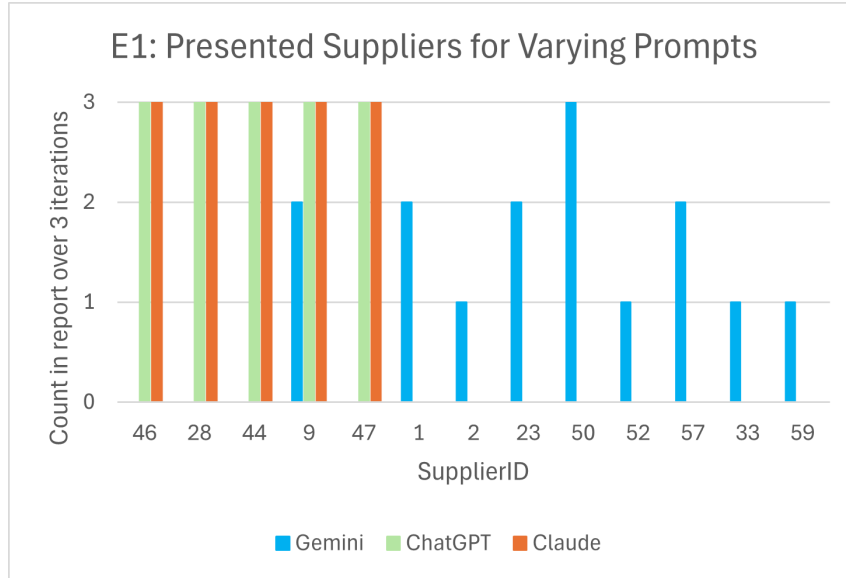


Figure 6.10: *The frequency of presented suppliers by Gemini, ChatGPT and Claude for T5, E1.*

Both Claude and Gemini are consistent in both environments, presenting the same five suppliers: 46, 28, 44, 9, and 47, as output to each of the modified prompts. Out of these five suppliers, only supplier 9 is highlighted by Gemini, and it is this in two out of three iterations. The only supplier highlighted by Gemini through all three prompts is supplier 50, which is not highlighted by Claude or ChatGPT. In total, over the three different iterations of prompt modifications, nine different suppliers are highlighted by Gemini. When shifting to E2, Figure 6.11 below illustrates the number of presented products for each prompt and genAI model.

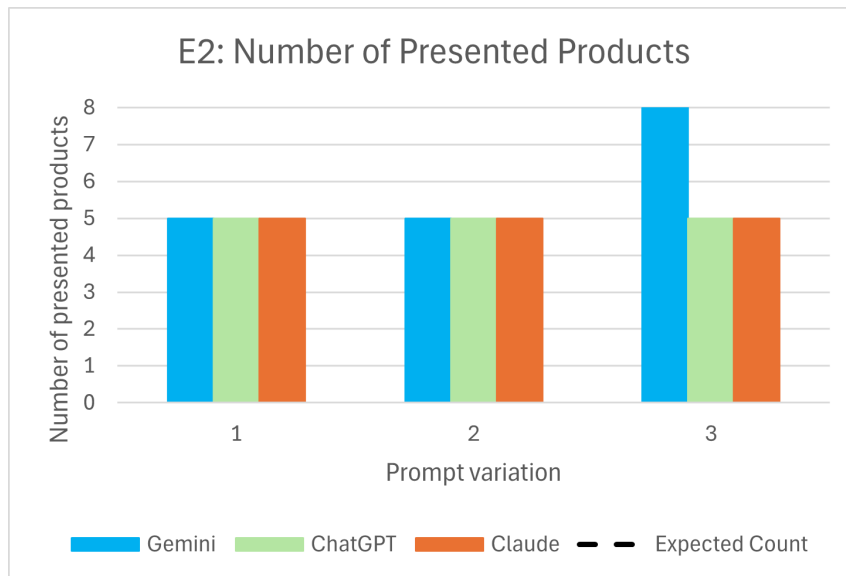


Figure 6.11: *The number of presented products by Gemini, ChatGPT and Claude for each iteration with a modified prompt in T5, E2, along with the expected number.*

As seen in 6.11, Claude and ChatGPT are again consistent in their outputs, providing five products each iteration. Gemini on the other hand are also presenting five products when prompted with the first two modified prompts, but when prompted with the third it provides eight products. Again, what products are presented differ, as illustrated below in Figure 6.12.

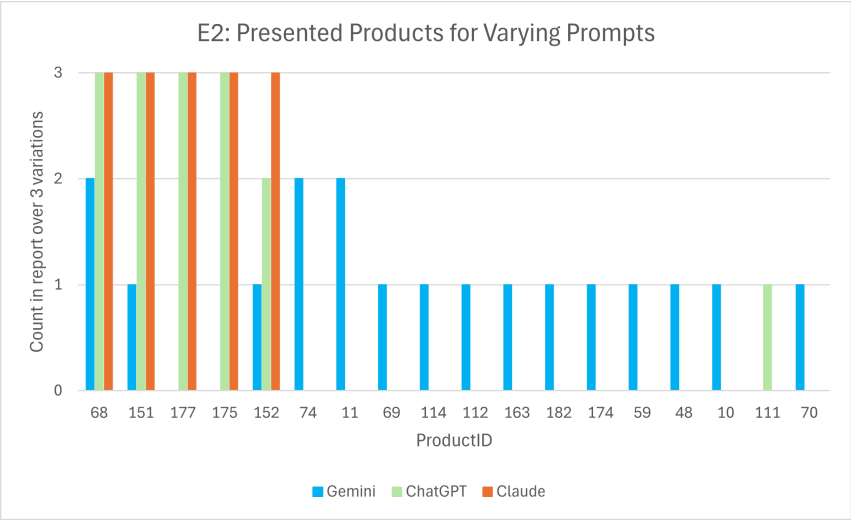


Figure 6.12: *The presented products by Gemini, ChatGPT and Claude during T5, E2 and their frequency of being deemed one of the five most problematic issues.*

Similarly to in E1, Figure 6.12 shows Claude presenting the same five products: product 68, 151, 177, 175, and 152, independently of the prompt version. ChatGPT is consistently presenting the same products as Claude when using prompt modification 1 and 3, but when using prompt modification 2 product 152 is replaced by product 111. Finally, Gemini presents 15 different products, out of which only three are the same as those presented by Claude and ChatGPT. Only products 68, 74, and 11 are represented more than once in the outputs provided by Gemini, where only 68 is presented by the other two models as well.

To summarize all test results for Gemini, ChatGPT and Claude, Figure 6.13 displays all test results for the three models in E1. A line being closer to the edge indicates a good performance from the model in that corresponding metric. The figure displays both test environments separately. Metrics which are not presented through values between zero and one have been normalized to display all tests a similar scale.

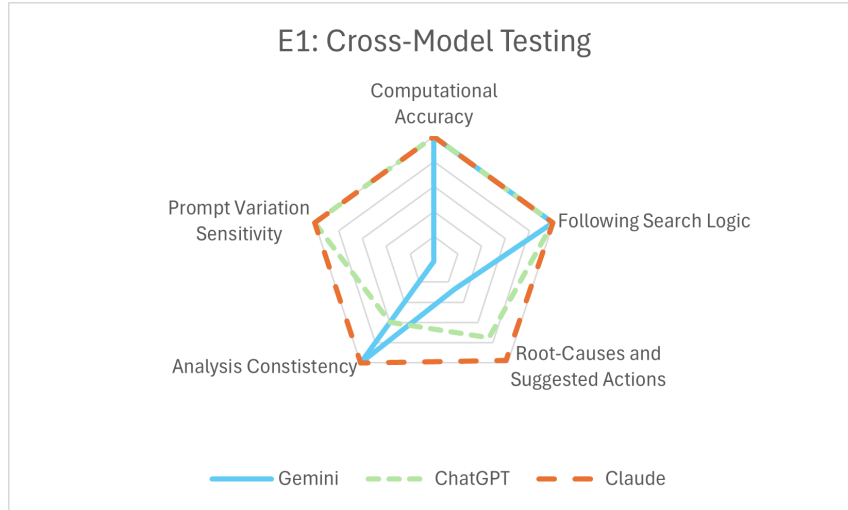


Figure 6.13: *Cross-model comparison displaying all test results from applying the SAFE-PE framework to the adjusted prompt-framework in E1*

Figure 6.13 displays Claude having the best overall performance, followed by ChatGPT. The difference between the two models is connected to analysis consistency, where ChatGPT's performance drops. Gemini proves efficient regarding CA, following search logic and being consistent in its analysis. However, it showed extreme sensitivity to smaller variations in the prompt, yielding a poor performance in that aspect.

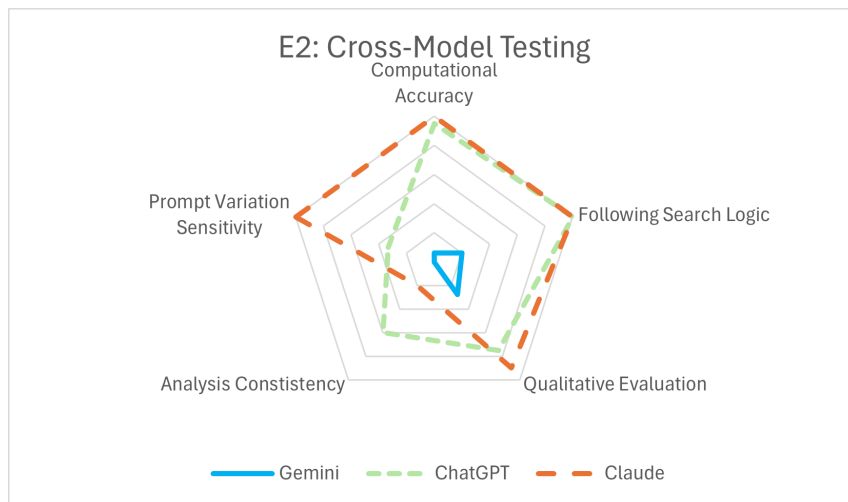


Figure 6.14: *Cross-model comparison displaying all test results from applying the SAFE-PE framework to the adjusted prompt-framework in E2*

Figure 6.14 shows a similar hierarchy in E2 as in E1 between the three models, with Claude yielding higher overall metrics compared to the others. Claude does show one weakness in this case, where it fails in being consistent during T4. ChatGPT proves efficient in CA and following search logic, but lacks performance in regards to the

qualitative evaluation, its consistency of analysis, and being sensitive to variations in the prompt. Gemini experiences poor performance across all metrics.

Overall, no test value is better for E2 compared to E1. Claude and ChatGPT retains high computational accuracy and are able to follow an established search logic to find deviations in both test environments. The results are further discussed in the next chapter.

7 Discussion

In this chapter, test results from applying the SAFE-PE framework to the prompts used to analyze PipeChain SCM's inbound process are discussed and analyzed. The chapter ends with an overall discussion regarding opportunities and challenges.

7.1 Phase 1 - Quantitative Metrics

Looking at the quantitative aspects of generated output, emphasis is put on reliability. GenAI being able to produce and present numerical values of high credibility is crucial in systems supported by genAI analysis.

T1: Calculation Accuracy

Claude, ChatGPT and Gemini all show high computational accuracy in test environment 1. For E1, all presented values are computed correctly, indicating the prompt does provide enough detail for how KPIs are computed from the dataset and how the model should analyze them. It also indicates the models' ability to perform numerical analysis on supply chain data, and their ability to structure data on different levels. In E1, MS Copilot is the exception, reaching only a 25 percent accuracy even though it only presents eight values in total. This indicates that this model is not developed for this kind of analysis, but rather textual problems or direct questions or tasks not demanding deeper analysis.

When observing the results from E2, it is clear that the models differ in performance. Claude is again faultless, providing a 100 percent score. ChatGPT on the other hand experiences some inaccuracies. Here, the inaccuracies regard the interpretation of what a forecast deviation is, as ChatGPT classified empty rows as a deviation when these instead should have been ignored as they indicate the lack of forecast, not a poor performing one. This was however not explicitly specified in the prompt. Gemini performs drastically worse than ChatGPT. Here the low score can be derived from the failure of identifying correct deviations overall, and thus it presents products that do not face issues as critical. Looking at what calculations Gemini found more difficult, no distinct pattern can be identified. This indicates the issue being connected to Gemini's interpretation of the prompt, rather than it finding certain calculations more difficult. MS Copilot fails completely here, having only presented one value and this being incorrect. Thus, as presented in Section 6.2 this model was excluded from further testing and analysis.

The trend that most models perform worse in E2 than E1 regarding CA may suggest that the prompt was lacking in clarity and directions, and that the more complex the analysis to be made is, the more explicit prompt is required to avoid that the genAI draws the wrong conclusions.

Focusing instead on the number of generated values, and excluding MS Copilot from the analysis, Gemini can be extinguished as the model providing the least

number of digits. This may suggest that this model does not go as deep in its analysis as the other two that provide approximately three times more values in both environments. ChatGPT is the model that provides the most values, yet, as previously stated, it does not achieve a perfect CA-score. Moreover, the difference in the number of generated values between Claude and ChatGPT is smaller in E1 than E2. In E1, ChatGPT provides 10,6 percent more values than Claude, whilst in E2 the number is 97,5 percent. The number of presented values being similar in E1 and then differ largely in E2 could suggest that the prompt for E2 leaves more room for interpretation by the genAI models, further strengthening the theory that the more complex the analysis is, the more guidance in the prompt is needed.

T2: Deviation Prioritization Accuracy

All genAI models perform well in identifying correct issues when including explicit instructions of how to prioritize deviations in the prompt in E1. This suggests that instructing genAI in a detailed manner will point its analysis in a certain direction. However, to exactly describe how suppliers or products should be selected and presented entails a more static analysis, leaving less room for genAI to adapt its visualization to what insights it finds in the data.

In E2 however, Gemini drastically drops to 20 percent whilst both Claude and ChatGPT still perform faultlessly. In this environment, Gemini only manages to correctly identify the product experiencing the most issues, whilst the other four products presented are not in the top five as instructed. The second presented product is the sixth most crucial product, but the other three presented products seem to have been randomly selected. When looking at the top five products presented from T1 in E2 by Gemini, none of them are correctly identified as the top five most critical products, suggesting that even though the results in T2 are undesirable, they have improved by the more explicit instructions and guidance in the prompt. This suggests that Gemini does not have analysis capabilities to the same extent as Claude and ChatGPT, as the analysis described in E2 seem to be too complex.

As ChatGPT in T1 experienced some difficulties with E2, presumably because of the more complex nature of the analysis, the results from this test strengthens the theory that a stricter prompt is required when the analysis becomes more complex. This is further supported by the improvement of Gemini's performance as well.

7.2 Phase 2 - Qualitative Metrics

In regards to the qualitative aspects of generated output, emphasis is put on the interpretation of the process and the clarity of generated text. GenAI being able to produce and present clear and well motivated root-causes and action plans is crucial in systems supported by genAI analysis.

T3: Expert Opinions

As presented in Section 6.2.2, Gemini achieved the lowest score among the evaluated models across all four assessed aspects, including the lowest possible score in both

root-cause suggestions and suggested actions in the test environments. The consistently overall low performance suggests limitations in the model's ability to reason within the operational context and to synthesize meaningful process-related insights. When in E2, in four out of five evaluated products, Gemini identified the absence of forecasting as the primary root-cause and recommended establishing forecasts as the corrective action. In E1, all presented root-causes and recommended actions were the same for all suppliers. This repetitive pattern indicates a tendency toward generic recommendations rather than context-specific analysis, which may suggest insufficient domain adaptation or limited contextual understanding of the provided process data. Furthermore, the findings imply that the prompt context alone may not have been sufficient for the model to generate operationally relevant insight.

In contrast, both Claude and ChatGPT demonstrated substantially stronger performance across the evaluated aspects. Although the models achieved identical scores in root-cause suggestions in E2, Claude outperformed ChatGPT in all remaining categories no matter the environment, indicating a greater ability to interpret process-related information and generate value-adding qualitative insights. In particular, Claude achieved the maximum score in both process interpretation and customer value, suggesting a strong capability to understand and communicate operational issues in a meaningful manner.

Despite these strong results, Claude received comparatively lower scores in suggested actions and, in E2, root-cause analysis, indicating that accurate process interpretation does not necessarily translate into actionable operational recommendations. This may suggest that additional contextual information, such as process-specific constraints, organizational objectives, or historical operational data, is required for genAI models to generate more practically applicable recommendations. Furthermore, the results indicate that the evaluated models may currently be more effective as analytical support tools rather than autonomous decision-making systems.

7.3 Phase 3 - Robustness Metrics

Looking at the robustness aspects of generated output, emphasis is put on reproducibility and stability. GenAI being able to produce and present output of high credibility is crucial in systems supported by genAI analysis.

T4: Consistency of Deviation Reliability

Observing the results from T4, it is clear that there is a difference in performance of the genAI models between the different environments. The CDR values are consistently higher in E1 compared to E2, indicating that, as previously discussed, the models generally demonstrate stronger data analysis capabilities under the conditions defined in E1. In E1, the analysis is conducted on one KPI alone; DP, whilst E2 includes two KPIs; SL and FA. Therefore, the results from T4 suggest that the genAI models tested in general perform a more consistent analysis when instructed to assess one KPI at a time, and that contextual or structural differences between the environments significantly influence the model output quality. Notably, both Claude and Gemini achieve a maximum score of 1 in E1, while their performance

drops in E2, highlighting a potential sensitivity to differences between test environments. This is further supported by the results from T2, where all tested models achieved perfect test scores in E1, and only Gemini not finding the correct set of products for E2.

ChatGPT is the only tested model that presents the desired number of deviations in both test environments. In E2, Claude and Gemini both have iterations where they present four or six deviations, despite the prompt explicitly instructing the models to present five issues. This further strengthens the theory of genAI not being able to interpret more complex data on its own, but rather, in those cases, need more detailed instructions on how to perform its analysis. The number of issues presented, however, does not have to be connected to how the genAI performs its analysis. An incorrect search logic for issues should still lead to five presented deviations, which might indicate these inconsistencies being caused by hallucinations in Claude and Gemini's cases.

Moreover, in E1 the total number of highlighted suppliers across all genAI models is 8, which corresponds to approximately 12 percent of all listed suppliers. Five suppliers corresponds to 7,6 percent. In E2 on the other hand, a total of 27 products were emphasized as critical across all three models. This corresponds to approximately 14 percent of all products available in the data. However, five products correspond to only 2,6 percent, suggesting that, in general, the models are better at confidently identifying issues in E1 than E2.

Looking at the results from a different perspective though, the spread of emphasized suppliers and products could also suggest that the AI model is conducting its analysis in different ways and thus finding different kinds of problems which also is the desired output. Going back to the analysis from T2, it would be relatively easy to code an algorithm to find certain problems in data the way a human would, but what is desired when using genAI is to get new perspectives and analysis, with the objective of finding problems and issues otherwise ignored. However, when differentiating the method of analysis, transparency becomes even more crucial as the user needs to understand the conducted analysis and reasoning behind the output provided.

T5: Sensitivity Testing

As the primary focus of this study is the data analysis capability of the genAI models, the modifications of the prompt were restricted to the first three components: purpose, context, and explanation of input data. These components form the foundation for how the models conduct the analysis, in accordance with the steps outlined in the task breakdown.

ChatGPT scores a perfect score in T5 in E1, suggesting that prompt modification does not affect the analytical accuracy if aligned well with the model's strengths. However, this improvement is not consistent in E2, where performance drops drastically. This may further indicate that the more complex the analysis, the more important is the prompt and its structure. Similarly, Gemini fails completely in T5, indicating that prompt changes may negatively affect performance if they do not match the model's interpretation patterns. The performance of Gemini in T5 is

particularly striking and indicates that even minor adjustments to prompt structure can significantly disrupt the model’s ability to interpret and analyze data correctly. This raises some concern regarding the fragility of these models when exposed to variations in input formulations. However, as this is not the case for Claude or ChatGPT it may rather suggest that the problem lies with the genAI model and how it has been trained, rather than the structure of the prompt. As this study did not include fine tuning, the results solely rely on the pre-training of the AI-models. This test may have highlighted that Gemini has not been as trained on data analysis as Claude and ChatGPT, or at least trained to perform analysis in a different way.

The absence of a clear pattern across models does, however, complicate the interpretation of the results. While some models handle prompt modifications well in specific cases, others deteriorate, suggesting that prompt effectiveness is highly model-dependent and lacks transferability. This limits the practical applicability of prompt engineering as a standardized method for this kind of analysis. Additionally, the relatively small number of iterations and prompt variations may restrict the generalizability of the findings, meaning that the observed effects could be partly driven by randomness rather than systematic differences. This goes for all tests.

7.4 Implications of the Evaluation

Continuing to follow the DSR framework, this chapter addresses the fifth step: evaluation. Overall, the findings suggest that while prompt engineering can be used for data analysis task performed by genAI, its effectiveness is highly dependent on both the model and the testing environment. This highlights the importance of aligning prompt design not only with the task structure, but also with the specific characteristics of the model and its operating context.

A notable difference is found between the results from Claude in T4 and T5. In T5, Claude performs faultlessly, scoring 1 in both environments. However, in T4 there is, as previously mentioned, a drastic drop in performance between the two environments indicating worse performance in this test. A reason for this difference may be the nature of the tests; in T5, the prompts are slightly modified each iteration, providing the genAI, in a way, with a new task each iteration, whilst in T4, the prompt remains the same each iteration. This may lead to the genAI believing that the previous answer was not satisfactory and therefore tries to find another way of analysis, resulting in a different output than the previous iteration. Thus, the results does not necessarily suggest that Claude lacks in analysis capabilities, but rather that it is adapting its analysis to the user.

Common for all genAI models are that they perform equivalent or worse in E2 than in E1. As touched upon previously, this may indicate a lack of analysis capabilities when the analysis becomes more complex and involves more than one aspect to consider. This may be supplemented by improving the prompt, making it even more explicit how the analysis is to be performed and what to prioritize. This is supported by the results in T2, where more explicit instructions resulted in an improvement for both ChatGPT and Gemini. This is, however, made more complex by the desire to let the genAI make its own analysis with the objective of being

bias free and not too restricted by the human mind. Another way to address this is through fine tuning. By further training the AI-model on context specific data, its analysis capabilities may increase with that domain, resulting in a better root-cause analysis and recommended actions.

A key opportunity following the automation of the calculation and interpretation of KPIs is the reduction of manual labor for filtering and analyzing. Previous to the assistance of genAI, the user had to either use the filters directly in PC, or export the data to Excel and then in Excel manually filter or use other methods to find the potential products or suppliers facing issues or poor trends. Once these have been identified, the user must analyze the patterns and find possible root causes to the issues as well as how to handle them. This way of working is time consuming and may be prone to human error. If found to be accurate and trustworthy, implementing a genAI model in this process can save much time, and the KPIs can be visualized to make it easier for the user to understand the trends and what to do.

Regarding the qualitative aspect of the output generated by genAI, a notable observation from the results is that the evaluated models generally achieved higher scores in process interpretation than in suggested actions. This may indicate that current genAI models are more capable of identifying and describing operational issues than proposing practically feasible improvement initiatives. Consequently, the findings suggest that genAI may currently provide the greatest value as a supporting analytical tool in early-stage process analysis, while human expertise remains necessary for evaluating and implementing operational decisions. However, this is also an aspect that may be improved through fine tuning.

Finally, this thesis puts emphasis on genAI's ability to analyze and present insights from data, rather than how an actual integration of supported decision-making in organizations can be executed. Hence, the human aspects of AI-driven decision support presents an opportunity. Topics such as trust from decision-makers, applicability on different organizational levels, and AI making independent decisions are all left for further investigation.

8 Conclusion

In this chapter, the purpose of the thesis and its research questions are revisited and answered based on the findings. Then, the research contribution is stated after which suggestions for future research are discussed, highlighting areas where further analysis have potential to generate even more insights.

8.1 Addressing the Purpose and Research Questions

As stated in Chapter 1, the purpose of this thesis was to explore how genAI could be integrated into digital supply chain environments to create value for system users. This was addressed with the help of the case company PC SCM and their digital landscape PC, used by supply chain partners for information sharing and tracking of data. By investigating how genAI can support operational decision-making through automating the interpretation of KPIs, the study aimed to simplify and facilitate the identification of issues through the provision of improved decision support. For this purpose, three research questions were formulated.

8.1.1 RQ1

How can genAI-driven analysis be used to interpret changes in KPIs and in this way identify real time issues?

In today's digital landscape, several of the big genAI companies offer a module for analyzing data when using their model. This thesis evaluated how well ChatGPT, Claude, Copilot and Gemini interpret, structure and present data in a real supply chain setting. Results indicate a dependency of what model is being used for the analysis, rather than the strengths and weaknesses of the technology itself. Well-performing models are able to reliably interpret and present data correctly, especially in less complex environments or when explicitly told how to perform calculations.

All models show less robustness in environments with a higher complexity, at least in the case where genAI is supposed to find its own way of finding deviations. Therefore the technology's analysis function is more suitable to implement in environments with a lower complexity. Using the tool in an environment with more parameters to take into account demands a more detailed task breakdown in the prompt, as well as clear instructions of how the analysis should be executed.

Furthermore, multiple genAI models proved sufficient in looking at historical data and pointing out issues in a given setting. Out of the four different models, three managed to create their own search logic and present problematic suppliers and products according to that logic.

In summary, the study found that through prompt engineering, genAI models can be used to highlight and present changes in KPIs in data handed to it, as well as short root-cause analysis alongside recommendation actions in a much less time consuming manner than if done manually by humans. If done with real-time data,

this will thus facilitate the real-time risk handling.

8.1.2 RQ2

Can genAI be used to prioritize and present KPIs based on operational context and decision-making needs?

This research question was investigated through several tests in which the AI model was given varying degrees of freedom in selecting analysis methods and prioritization criteria. The results indicate that AI can be used to prioritize and present KPIs based on operational context and decision-making needs. Across experiments, the models were generally capable of identifying and highlighting the most critical operational issues, even when only limited instructions were provided. However, the consistency and accuracy of the prioritization improved significantly when clearer guidance and contextual constraints were introduced.

The findings suggest that there is a tradeoff between AI autonomy and analytical reliability. Providing more structured instructions reduced the variability in the output and increased the alignment with expected output. At the same time, overly restrictive prompting risks limiting the AI's ability to identify alternative perspectives, unexpected relationships, or less obvious performance issues that could still be relevant for decision-making.

This highlights the importance of balancing flexibility and control when implementing AI-supported KPI analysis. Rather than replacing managerial judgment, the AI appears most valuable as a decision-support tool capable of filtering large amounts of operational data and presenting prioritized insights in a structured manner. The results therefore indicate that AI has strong potential to support context-aware KPI prioritization, but that the quality of the output remains highly dependent on prompt design, contextual information, and human oversight.

8.1.3 RQ3

To what extent can genAI find causes of shifts in KPI-values and give recommended actions?

Due to the limitations in T3 regarding the expert assessments that being conducted by researchers and not industry experts, it is not possible to address this research question definitively. Therefore, findings regarding genAI's ability to produce qualitative insights should be viewed as indications of the technology's potential, and will need further confirmation by future research.

However, the findings could indicate that genAI's ability to identify causes behind shifts in KPI values is highly dependent on the model conducting the analysis. While some models struggled to provide relevant explanations and actionable recommendations, others demonstrated a greater capability to generate coherent, motivated, and contextually grounded suggestions. Nevertheless, none of the evaluated models achieved consistently accurate results across all cases, highlighting limitations in the current application of genAI for root-cause analysis within operational environments.

Two potential explanations for these limitations can be identified. First, the prompts

used in this thesis may not have provided sufficient contextual specificity regarding the operational processes, KPI interdependencies, and decision-making constraints. This suggests that more advanced prompt engineering, including structured contextual input and clearer analytical instructions, could improve the relevance and precision of the generated outputs. Second, the results indicate that fine-tuning or domain-specific training may be necessary to enable genAI models to better understand case-specific relationships and operational dynamics.

Although the results do not allow for a definitive conclusion regarding the extent to which genAI can be used to give recommended actions to shifts in KPIs, they clearly demonstrate that genAI performance is strongly influenced by the choice of model, and how this has been pre-trained. Consequently, the study suggests that genAI should not be viewed as a standalone autonomous decision-making system, but rather as a decision-support tool whose effectiveness depends heavily on the quality and structure of the information provided.

8.2 Research contribution

By developing a prompt framework and applying it within the operational environment of PC, this study provides a practical example of how genAI can be utilized in a digital supply chain context for operational data analysis and decision support. The study thereby demonstrates how LLMs can be integrated into existing supply chain systems to interpret, prioritize, and present KPI-related insights.

Furthermore, the thesis contributes to the academic understanding of AI within SCM by exploring the use of genAI in analytical tasks traditionally performed by human analysts. While previous research has largely focused on forecasting, automation, or optimization, this study examines the role of genAI in contextual KPI analysis and decision-support processes, an area that remains relatively unexplored.

In addition, the research contributes methodologically through the development and evaluation of a structured prompt framework. The findings demonstrate how prompt design, contextual information, and the level of AI guidance significantly influence the analytical quality and consistency. This provides a foundation for future studies investigating prompt engineering as a method for improving AI-supported business analysis. Furthermore, the thesis provides a framework designed to analyze and present data that, regardless the industry or setting, can be applicable to automated data analysis, or further developed for future research.

From a practical perspective, this study highlights both the opportunities and limitations of using genAI as a support tool for operational decision-making. The results indicate that genAI can be used to reduce the effort required to analyze large amounts of unstructured operational data while also identifying critical issues and prioritizing KPIs. At the same time, the study identifies challenges related to reliability, consistency, and the balance between AI autonomy and human oversight, providing guidance for organizations seeking to implement similar solutions.

8.3 Limitations and Suggestions for Future Research

Building on this study, there are multiple areas of interest for future research, many driven by the limitations of this study.

When conducting this study, the latency time has been deliberately excluded from the evaluation, with the sole emphasis placed on analytical capability. While this simplification aligns with the specific use case examined, where analysis is conducted at most once per day, it restricts the applicability of the findings. In more time-sensitive contexts, where data requires more frequent analysis, latency may become a decisive factor influencing model suitability. Consequently, the exclusion of this dimension limits the generalizability of the results and risks overlooking important trade-offs between analytical performance and response efficiency. Future research should therefore incorporate latency as a key evaluation metric to provide a more comprehensive assessment of prompt performance.

Similarly, the computational resources required by each model have not been considered, representing another significant limitation. Differences in computational demand can have substantial implications for scalability, cost-efficiency, and practical implementation, particularly in industrial settings. Ignoring this aspect may lead to an incomplete understanding of the feasibility of deploying these models in real-world applications where resource constraints are often critical. While the focus of this thesis has been accuracy, prioritization, and consistency, future research should rather put emphasis on latency, cost, and scalability.

Beyond technical performance, an important topic for future research lies in examining the impact of genAI on decision-making processes. While this study evaluates the prompt's and the model's ability to generate analytical outputs, it does not address how these outputs influence user behavior or decision quality. Investigating how system users interact with and rely on genAI-generated analyses would provide valuable insight into whether such tools genuinely simplify decision-making or merely shift complexity elsewhere. A particularly relevant approach would be to compare decisions made with the support of genAI-generated reports to those based solely on human judgment. Such a comparison could reveal not only differences in decision quality but also changes in work practices, trust, and cognitive load, thereby offering a more holistic evaluation of the practical value of genAI in this context.

Regarding the results, the evaluation approach in T3 introduces certain limitations. It is important to acknowledge that since the assessments were conducted by the researchers rather than external domain experts, the results may be influenced by subjective interpretation and potential researcher bias. Although the evaluations were conducted independently, the absence of practitioner-based assessments may reduce the external validity of the findings. Future research could therefore benefit from involving industry experts in the evaluation process to obtain a more practically grounded assessment of the generated outputs.

Moreover, it is important to acknowledge the small sample size for all test. Here, a trade-off had to be made between testing broadly or more detailed due to time

constraints. The decision to test more broadly was made, with the motivation that it is of more interest at start, to get an overall indication of the capabilities and possibilities, rather than an in-depth statistically supported understanding of a single aspect. This would, on the contrary, be of more interest in further research.

Lastly, the thesis deploys a single case study, and thus the generalizability is restricted. Despite looking at data collected in different supply chain contexts, the study only covers PC SCM's inbound process, and is therefore not necessarily directly applicable to other digital supply chain environments. However, although the findings are context-dependent, the investigated use case addresses challenges that are prevalent in many supply chain environments, providing insights that may be transferable to similar organizational settings. To increase the generalizability of the findings, future research should include a greater variation of industries and systems.

References

- Aaijaz, N., Deshmukh, B. S., Kumar, M., Sheth, N., Mishra, P., and Jaiswal, V. K. (2025). “Revolutionizing Enterprise Decision-Making: Leveraging AI for Strategic Efficiency and Agility”. In: *Advances in Consumer Research* 2.4, pp. 3464–3470.
- Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., Bodenstein, S.W., Evans, D.A., Hung, C.-C., O’Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A.I., Cowie, A., Figurnov, M., Fuchs, F.B., Gladman, H., Jain, R., Khan, Y.A., Low, C.M.R., Perlin, K., Potapenko, A., and, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E.D., Zielinski, M., Židek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D., and Jumper, J.M. (2024). “Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3”. In: *Nature* 630, pp. 493–500. DOI: 10.1038/s41586-024-07487-w.
- Association for Supply Chain Management (2022). *Supply Chain Operations Reference Model – SCOR*. Accessed: 2026-04-13. URL: <https://scor.ascm.org/processes/introduction>.
- Athaluri, S. A., Manthana, S. V., Kesapragada, V. K. M., Yarlagadda, V., Dave, T., and Duddumpudi, R. T. S. (2023). “Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing through ChatGPT References”. In: *Cureus* 15.4.
- Baah, C., Opoku Agyeman, D., Acquah, I. S. K., Agyabeng-Mensah, Y., Afum, E., Issau, K., Ofori, D., and Faibil, D. (2022). “Effect of Information Sharing in Supply Chains: Understanding the Roles of Supply Chain Visibility, Agility and Collaboration on Supply Chain Performance”. In: *Benchmarking: An International Journal* 29.2, pp. 434–455.
- Bahoo, S., Cucculelli, M., and Qamar, D. (2023). “Artificial Intelligence and Corporate Innovation: A Review and Research Agenda”. In: *Technological Forecasting and Social Change* 188. DOI: 10.1016/j.techfore.2022.122264.
- Bahroun, Z., Saihi, A., As’ad, R., and Tanash, M. (2025). “A Systematic Analysis of Generative Artificial Intelligence for Supply Chain Transformation”. In: *Supply Chain Analytics* 13, p. 100188. DOI: 10.1016/j.sca.2025.100188.
- Balfaqih, H., Nopiah, Z. Mohd., Saibani, N., and Al-Nory, M. T. (2016). “Review of Supply Chain Performance Measurement Systems: 1998–2015”. In: *Computers in Industry* 82, pp. 135–150. DOI: 10.1016/j.compind.2016.07.002.
- Booyse, D. and Scheepers, C. B. (2024). “Barriers to adopting automated organisational decision-making through the use of artificial intelligence”. In: *Management Research Review* 47.1, pp. 64–85. DOI: 10.1108/MRR-09-2021-0701.

- Brocke, J. vom, Hevner, A., and Maedche, A. (Sept. 2020). "Introduction to Design Science Research". In: pp. 1–13. ISBN: 978-3-030-46780-7. DOI: 10.1007/978-3-030-46781-4_1.
- Buchari, M. A. and Tontowi, A. E. (2024). "Design of Supply Chain Performance Measurement Using A Hybrid of SCOR Model and Multi Criteria Decision Making Method in the Seaweed Industry". In: *2024 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pp. 604–608. DOI: 10.1109/IEEM62345.2024.10857222.
- Bueno, A., Filho, M. G., and Frank, A. G. (2020). "Smart Production Planning and Control in the Industry 4.0 Context: A Systematic Literature Review". In: *Computers & Industrial Engineering* 149, p. 106774.
- Cai, J., Liu, X., Xiao, Z., and Liu, J. (2009). "Improving Supply Chain Performance Management: A Systematic Approach to Analyzing Iterative KPI Accomplishment". In: *Decision Support Systems* 46.2, pp. 512–521. DOI: 10.1016/j.dss.2008.09.004.
- Castelvecchi, D. (2016). *Can We Open the Black Box of AI?* Accessed 27 January 2026. URL: <https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731>.
- Çelikmasat, G., Özgövde, A., and Aydemir, F. B. (2025). "Generating Business Process Models with Open Source Large Language Models Using Instruction Tuning". In: *Lecture Notes in Computer Science*. Springer, pp. 269–284. DOI: 10.1007/978-3-032-12089-2_17.
- Dabbous, A., Merhej Sayegh, M., Aoun Barakat, K., and Veglianti, E. (2025). "Corporate Artificial Intelligence Readiness and Its Contribution to Supply Chain Resilience". In: *International Journal of Production Research*, pp. 1–23. DOI: 10.1080/00207543.2025.2578299.
- Danilevskiy, M., Perez-Tellez, F., and Buscaldi, D. (2025). "Implementing ethical principles in AI: an initial discussion". In: *AI Ethics* 5, pp. 3549–3555. DOI: 10.1007/s43681-025-00710-y.
- Dash, R., McMurtrey, M., Rebman, C., and Kar, U. K. (2019). "Application of Artificial Intelligence in Automation of Supply Chain Management". In: *Journal of Strategic Innovation and Sustainability* 14.3, pp. 43–53. DOI: 10.33423/jsis.v14i3.2105.
- Denscombe, M. (2010). *The Good Research Guide: for Small-Scale Social Research Projects*. 4th ed. Maidenhead: Open University Press.

- Di Mauro, C., Peters, E., and Carnovale, S. (2024). “A Seat at the Table: The Future of Purchasing and Supply Management”. In: *Journal of Purchasing and Supply Management* 30.1, p. 100908.
- Dumitrascu, O., Dumitrascu, M., and Dobrotă, D. (2020). “Performance evaluation for a sustainable supply chain management system in the automotive industry using artificial intelligence”. In: *Processes* 8.11, p. 1384. DOI: 10.3390/pr8111384.
- Emon, M. M. H., Khan, T., and Siam, S. A. J. (2024). “Quantifying the Influence of Supplier Relationship Management and Supply Chain Performance: An Investigation of Bangladesh’s Manufacturing and Service Sectors”. In: *Brazilian Journal of Operations & Production Management* 21.2, pp. 2015–2015.
- European Commission (Jan. 2021). *Industry 5.0: Towards more sustainable, resilient and human-centric industry*. Accessed: 2026-05-11. URL: https://research-and-innovation.ec.europa.eu/news/all-research-and-innovation-news/industry-50-towards-more-sustainable-resilient-and-human-centric-industry-2021-01-07_en.
- European Parliament and Council (2024). *Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence*. Accessed 4 February 2026. URL: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>.
- Fares, N., Anass, C., Shardeo, V., and Garza-Reyes, J. A. (2025). “Supply Chain Management in the Industry 5.0 Era: Strategic Implications”. In: *Benchmarking: An International Journal*. Ahead-of-print. DOI: 10.1108/BIJ-12-2024-1136.
- Fatawi, I., Asy’ari, M., Hunaepi, H., Samsuri, T., and Bilad, M. R. (2024). “Empowering Language Models Through Advanced Prompt Engineering: A Comprehensive Bibliometric Review”. In: *Indonesian Journal of Science and Technology* 9.2, pp. 441–462. DOI: 10.17509/ijost.v9i2.71481.
- Ferreira, A. C. A., Francisco, M. B., and Pinho, A. F. de (2025). “The Use of Artificial Intelligence in Supply Chain Management: Systematic Literature Review and Future Research Directions”. In: *IEEE Access* 13, pp. 157828–157841. DOI: 10.1109/ACCESS.2025.3603866.
- Genetti, S., Longobardi, A., and Iacca, G. (2025). “Evolutionary Reinforcement Learning for Interpretable Decision-Making in Supply Chain Management”. In: *Applications of Evolutionary Computation*. Cham: Springer. DOI: 10.1007/978-3-031-90062-4_12.
- Ghobakhloo, M., Iranmanesh, M., Foroughi, B., Tseng, M. L., Nikbin, D., and Khanfar, A. A. A. (2025). “Industry 4.0 Digital Transformation and Opportunities for Supply Chain Resilience”. In: *Production Planning & Control* 36.1, pp. 61–91. DOI: 10.1080/09537287.2023.2252376.

- Giray, L. (2023). "Prompt Engineering with ChatGPT: A Guide for Academic Writers". In: *Annals of Biomedical Engineering* 51, pp. 2629–2633. DOI: 10.1007/s10439-023-03272-4.
- Golicic, S. L., Davis, D. F., and McCarthy, T. M. (2005). "A Balanced Approach to Research in Supply Chain Management". In: *Research Methodologies in Supply Chain Management*. Ed. by Kotzab, H., Seuring, S., Müller, M., and Reiner, G. Physica-Verlag HD. DOI: 10.1007/3-7908-1636-1_2.
- Goutier, M., Hirt, R., Köhl, N., and Satzger, G. (2019). "Machine Learning in Artificial Intelligence: Towards a Common Understanding". In: *Proceedings of the Hawaii International Conference on System Sciences*, pp. 5236–5245.
- Gunasekaran, A. and Kobu, B. (2007). "Performance Measures and Metrics in Logistics and Supply Chain Management: A Review of Recent Literature (1995–2004) for Research and Applications". In: *International Journal of Production Research* 45.12, pp. 2819–2840. DOI: 10.1080/00207540600806513.
- Harmon, P. (2003). *An Introduction to the Supply Chain Council's SCOR Methodology*. Business Process Trends Whitepaper.
- Heinrich, K., Janiesch, C., and Zszech, P. (2021). "Machine Learning and Deep Learning". In: *Electronic Markets* 31, pp. 685–695. DOI: 10.1007/s12525-021-00475-2.
- Helo, P. and Hao, Y. (2021). "Artificial intelligence in operations management and supply chain management: an exploratory case study". In: *Production Planning & Control* 33.16, pp. 1573–1590. DOI: 10.1080/09537287.2021.1882690.
- Hollmann, R. L., Scavarda, L. F., and Thomé, A. M. T. (2015). "Collaborative Planning, Forecasting and Replenishment: A Literature Review". In: *International Journal of Productivity and Performance Management* 64.7, pp. 971–993.
- Holmberg, S. (2000). "A Systems Perspective on Supply Chain Measurements". In: *International Journal of Physical Distribution & Logistics Management* 30.10, pp. 847–868. DOI: 10.1108/09600030010351246.
- Höst, M., Regnell, B., and Runeson, P. (2006). *Att genomföra examensarbete*. Lund: Studentlitteratur.
- Huang, K., Wang, K., Lee, P. K. C., and Yeung, A. C. L. (2023). "The Impact of Industry 4.0 on Supply Chain Capability and Supply Chain Resilience: A Dynamic Resource-Based View". In: *International Journal of Production Economics* 262, p. 108913. DOI: 10.1016/j.ijpe.2023.108913.

- Huang, W., Li, H., Qin, F., Wu, D., and Cheng K. abd Chen, H. (2026). “A prompt-guided generative language model for unifying visual neural decoding”. In: *International Journal of Neural Systems*. DOI: 10.1142/S0129065725500686.
- Huckenbeck, J. and Zatrochová, M. (2024). “Use of Artificial Intelligence in Supply Chain Management”. In: *Smart Cities: Importance of Management and Innovations for Sustainable Development*. Cham: Springer, pp. 299–309. DOI: 10.1007/978-3-031-56533-5_21.
- Ivanov, D. (2025). “Supply Chain Economics: Research Landscape and Future Directions”. In: *IFAC-PapersOnLine* 59.10. 11th IFAC Conference on Manufacturing Modelling, Management and Control MIM 2025, pp. 67–71. ISSN: 2405-8963. DOI: <https://doi.org/10.1016/j.ifacol.2025.09.013>.
- Jiang, C., Huang, S., and Shen, T. (2025). “Guiding the unseen: prompt-driven agency dynamics”. In: *Systems*. DOI: 10.3390/systems13070576.
- Jou, M. and Hao, Y. (2026). “Let technology serve learning”. In: *Interactive Learning Environments*. DOI: 10.1080/10494820.2026.2617983.
- Kache, F. and Seuring, S. (2017). “Challenges and Opportunities of Digital Information at the Intersection of Big Data Analytics and Supply Chain Management”. In: *International Journal of Operations & Production Management* 37.1, pp. 10–36.
- Kan, Z., Feng, L., Yin, Z., Qiao, L., Qiu, X., and Li, D. (2023). “A composable generative framework based on prompt learning for various information extraction tasks”. In: *IEEE Transactions on Big Data*. DOI: 10.1109/TBDATA.2023.3278977.
- Kembro, J., Näslund, D., and Olhager, J. (2017). “Information sharing across multiple supply chain tiers: A Delphi study on antecedents”. In: *International Journal of Production Economics* 193, pp. 77–86. ISSN: 0925-5273. DOI: <https://doi.org/10.1016/j.ijpe.2017.06.032>.
- Kembro, J., Näslund, D., and Olhager, J. (2025). “Information sharing in extended supply chains: why does it have to be so difficult?” In: *Production Planning & Control* 36.8, pp. 1029–1043. DOI: 10.1080/09537287.2024.2331742.
- Khokrale, R. (2025). “The Role of AI in Supply Chain Optimization: Enhancing Efficiency through Predictive Analytics”. In: *Journal of Procurement and Supply Chain Management* 4.2, pp. 55–75. DOI: 10.58425/jpscm.v4i2.442.
- Kim, S. K., Kim, G.-M., Cha, S., and Jung, M. (2025). “Verification of the Validity and Reliability of Therapeutic Communication Responses Generated by Large Language Models (LLMs): A Comparative Study of Prompt-Engineering Strategies”. In: *Journal of Korean Academy of Psychiatric and Mental Health Nursing* 34.Special Issue, pp. 23–35. DOI: 10.12934/jkpmhn.2025.34.s1.23.

- Kirshner, S (2024). “Artificial Agents in Operations Management Experiments”. In: *SSRN Electronic Journal*. UNSW Business School Research Paper Forthcoming. DOI: 10.2139/ssrn.4726933. URL: <https://ssrn.com/abstract=4726933>.
- Kuligin, L., Lammert, J., Ostapenko, A., Bressem, K., Boeker, M., and Tschochohei, M. (2025). “Prompt design for medical question answering with LLMs”. In: *Machine Learning with Applications*.
- Laeq, K. and Saleem, S. (2025). “SAFE-PE: A Systematic Assessment Framework for Evaluating Prompt Engineering in Generative AI”. In: *Sukkur IBA Journal of Computing and Mathematical Science* 9.2. Published online 16 January 2026. DOI: 10.30537/sjcms.v9i02.1776.
- Lakri, S. and Jemai, Z. (2015). “Performance Measurement and Management Systems of Supply Chains: A Review of the Challenges They Raise”. In: *Proceedings of the 2015 International Conference on Industrial Engineering and Systems Management (IESM)*. IEEE, pp. 920–929. DOI: 10.1109/IESM.2015.7380265.
- Larson, T., Bellah, J., and Du, F. (2023). “Conceptual meta-models: an example correlating Anthony’s triangle, Simon’s structure, and Stevens’ scale of measurement”. In: *The Journal of Applied Business and Economics* 25.3, pp. 128–136.
- Laskar, M.T.R., Bari, M.S., Rahman, M., Bhuiyan, M.A.H., Joty, S., and Huang, J.X. (2023). “A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets”. In: *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, pp. 431–469. DOI: 10.18653/v1/2023.findings-acl.29.
- Lee, E. J. and Park, S. J. (2025). “A Structured Prompt Framework for AI-Generated Biophilic Architectural Spaces”. In: *Journal of Building Engineering* 111, p. 113326. DOI: 10.1016/j.jobbe.2025.113326.
- Marinagi, C., Reklitis, P., Trivellas, P., and Sakas, D. (2023). “The impact of Industry 4.0 technologies on key performance indicators for a resilient Supply Chain 4.0”. In: *Sustainability* 15.6, p. 5185. DOI: 10.3390/su15065185.
- Maslaric, M., Petrovic, V., Fedorko, G., Nikolicic, S., and Mircetic, D. (2020). “Understanding the Importance of Purchasing Operations and Identification of Opportunities to Increase Their Sustainability Through Process Mapping”. In: *Prosperitas* 7.1, pp. 43–52.
- Mays, N. and Pope, C. (2000). “Qualitative research in health care: assessing quality in qualitative research”. In: *BMJ* 320.7226, pp. 50–52. DOI: 10.1136/bmj.320.7226.50.

- Meyer, F., Freitag, L., Hinrichsen, S., and Niggemann, O. (2024). “Potentials of Large Language Models for Generating Assembly Instructions”. In: *Proceedings of the 29th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*. Padua, Italy: IEEE. DOI: 10.1109/ETFA61755.2024.10710806.
- Mezmir, E. A. (2020). “Qualitative Data Analysis: An Overview of Data Reduction, Data Display, and Interpretation”. In: *Research on Humanities and Social Sciences* 10.21, pp. 15–27.
- Min, S., Zacharia, Z. G., and Smith, C. D. (2019). “Defining Supply Chain Management: In the Past, Present, and Future”. In: *Journal of Business Logistics* 40.1, pp. 44–55.
- Moore, O. (2026). *The Top 100 Gen AI Consumer Apps — 6th Edition*. <https://a16z.com/100-gen-ai-apps-6/>. Andreessen Horowitz, accessed 2026-04-23.
- Mulongo, N. Y. (2024). “Key Performance Indicators of Artificial Intelligence for IT Operations (AIOPS)”. In: *International Symposium on Networks, Computers and Communications (ISNCC)*. DOI: 10.1109/ISNCC62547.2024.10759030.
- Nazari-Ghanbarloo, V. (2020). “A Dynamic Performance Measurement System for Supply Chain Management”. In: *International Journal of Productivity and Performance Management*. DOI: 10.1108/ijppm-01-2020-0023.
- Novoszel, L. and Claus, I. (2024). “A methodological framework addressing challenges and opportunities in Supply Chain AI”. In: *IFAC-PapersOnLine* 58.19. 18th IFAC Symposium on Information Control Problems in Manufacturing INCOM 2024, pp. 349–354. ISSN: 2405-8963. DOI: <https://doi.org/10.1016/j.ifacol.2024.09.236>.
- Olavsrud, T. (2025). *10 most used gen AI tools in the enterprise*. <https://www.cio.com/article/3961674/10-most-used-gen-ai-tools-in-the-enterprise.html>. CIO, accessed 2026-04-23.
- Olhager, J. (2019). *Produktion och Logistik i försörjningskedjor*. 1st. Studentlitteratur.
- Özbiltekin-Pala, M. (2026). “Bibliometric Analysis of Performance Measurement in Digital Supply Chains”. In: *Ege Academic Review*. Advanced Online Publication, pp. 105–112.
- Pang, Y., Zhang, M., Liu, Y., Li, X., Wang, Y., Huan, Y., Liu, Z., Li, J., and Wang, D. (Nov. 2024). “Large language model-based optical network log analysis using LLaMA2 with instruction tuning”. In: *J. Opt. Commun. Netw.* 16.11, pp. 1116–1132. DOI: 10.1364/JOCN.527874.

- Phillips-Wren, G. and Håkansson, A. (2025). “Towards Using Prompt Engineering in Large Language Models to Assist Decision Making”. In: *Procedia Computer Science* 270. 29th International Conference on Knowledge-Based and Intelligent Information Engineering Systems (KES 2025), pp. 5225–5238. ISSN: 1877-0509. DOI: 10.1016/j.procs.2025.09.650.
- Phung, D-U., Phan, D. T-L., Do, N-H., and Le, C. T. D. (2024). “Upstream Supply Chain Visibility Measurement: A Case Study In Garment Industry”. In: *2024 International Conference on Logistics and Industrial Engineering (ICLIE)*, pp. 1–7. DOI: 10.1109/ICLIE61478.2024.11015392.
- PipeChain (n.d.). *Our History*. Accessed 2 Feb. 2026. URL: <https://www.pipechain.com/who-we-are/our-history/>.
- Ponce, P., Anthony, B., Bradley, R., Xu, W., and Molina, A. (2025). “A Comprehensive Framework Integrating ML, Automation Pyramid, and KPIs for Industry 5.0”. In: *Industry 5.0 Frameworks and Applications*, pp. 385–409. DOI: 10.1007/978-3-031-80760-2_25.
- Ramamoorthy, L. (2025). “Evaluating Generative AI: Challenges, Methods, and Future Directions”. In: *International Journal of Future Multidisciplinary Research* 7.1.
- Rasool, F., Greco, M., and Strazzullo, S. (2023). “Understanding the future KPI needs for digital supply chain”. In: *Supply Chain Forum* 25.4, pp. 550–561. DOI: 10.1080/16258312.2023.2253524.
- Rathkopf, C. (2026). *From Hallucination to Reliability: Generative Modeling and the Structure of Scientific Inference*.
- Riad, M., Naimi, M., and Okar, C. (2024). “Enhancing Supply Chain Resilience Through Artificial Intelligence: Developing a Comprehensive Conceptual Framework for AI Implementation and Supply Chain Optimization”. In: *Logistics* 8.4, p. 111. DOI: 10.3390/logistics8040111.
- Ronge, R., Maier, M., and Rathgeber, B. (2025). “Towards a Definition of Generative Artificial Intelligence”. In: *Philosophy & Technology* 38.1. DOI: 10.1007/s13347-025-00863-y.
- Rowley, J. and Slack, F. (2004). “Conducting a Literature Review”. In: *Management Research News* 27.6, pp. 31–39. DOI: 10.1108/01409170410784185.
- Roy, A., Sun, K., and Tobin, J. M. (2024). “Artificial Intelligence and Machine Learning: Definition of Terms and Current Concepts in Critical Care Research”. In: *Journal of Critical Care* 82. DOI: 10.1016/j.jcrc.2024.154792.

- Saenz, M. J. (2021). “Don’t overlook the importance of KPIs in AI/ML projects”. In: *Supply Chain Management Review*, pp. 8–9.
- Saunders, M., Lewis, P., and Thornhill, A. (2006). *Research methods for business students*. 4th ed. Pearson Education.
- Schoonenboom, J. (2023). “The Fundamental Difference Between Qualitative and Quantitative Data in Mixed Methods Research”. In: *Forum: Qualitative Social Research* 24.1, Article 11. DOI: 10.17169/fqs-24.1.3986.
- Shepherd, C. and Günter, H. (2006). “Measuring supply chain performance: current research and future direction”. In: *International Journal of Productivity and Performance Management* 55.3/4, pp. 242–258. DOI: 10.1108/17410400610653219.
- Singh, R. K. (2025). “Measuring Supply Chain Resilience Performance: Role of Data Analytics, Collaboration and Flexibility”. In: *Measuring Business Excellence* 29.1, pp. 121–136.
- Sreejesh, S., Mohapatra, S., and Anusree, M. R. (2013). *Business Research Methods: An Applied Orientation*. Springer Cham. ISBN: 978-3-319-00539-3. DOI: 10.1007/978-3-319-00539-3.
- Sunday Grève, S. (2025). “Turing’s Philosophy of AI”. In: *Journal for General Philosophy of Science* 56, pp. 513–529. DOI: 10.1007/s10838-024-09715-3.
- Triguero, I., Molina, D., Poyatos, J., Del Ser, J., and Herrera, F. (2023). “General Purpose Artificial Intelligence Systems (GPAIS): Properties, Definition, Taxonomy, Societal Implications and Responsible Governance”. In: *Information Fusion* 103, p. 102135. DOI: 10.1016/j.inffus.2023.102135.
- Tukamuhabwa, B. R., Stevenson, M., Busby, J., and Zorzini, M. (2015). “Supply chain resilience: definition, review and theoretical foundations for further study”. In: *International Journal of Production Research* 53.18, pp. 5592–5623.
- Van den Bogaert, J. and Van Jaarsveld, W. (2022). “Vendor-Managed Inventory in Practice: Understanding and Mitigating the Impact of Supplier Heterogeneity”. In: *International Journal of Production Research* 60.20, pp. 6087–6103.
- Varadejsatitwong, P., Banomyong, R., and Julagasigorn, P. (2022). “A Proposed Performance-Measurement System for Enabling Supply-Chain Strategies”. In: *Sustainability* 14.19. ISSN: 2071-1050. DOI: 10.3390/su141911797. URL: <https://www.mdpi.com/2071-1050/14/19/11797>.
- Venkatachalam, S. (2025). *Integrating Large Language Models with Network Optimization for Interactive and Explainable Supply Chain Planning: A Real-World Case Study*. DOI: 10.48550/arXiv.2508.21622.

- Verma, D., Singhal, R. K., Pandita, S., Garg, A., Singh, A., and M, A. N. (2025). “Exploring AI-Driven Supply Chain Innovations from Industry 4.0 to 6.0: A Comprehensive Review”. In: *2025 IEEE 7th International Conference on Computing, Communication and Automation (ICCCA)*, pp. 1–6. DOI: 10.1109/ICCCA66364.2025.11325432.
- Vilakati, S. (2025). “Prompt engineering for accurate statistical reasoning with large language models in medical research”. In: *Frontiers in Artificial Intelligence* 8. DOI: 10.3389/frai.2025.1658316.
- Voss, C., Tsikriktsis, N., and Frohlich, M. (2002). “Case research in operations management”. In: *International Journal of Operations & Production Management* 22.2, pp. 195–219. DOI: 10.1108/01443570210414329.
- Wang, S., Zhao, Y., Hou, X., and Wang, H. (2025). “Large Language Model Supply Chain: A Research Agenda”. In: *ACM Transactions on Software Engineering and Methodology* 34.5. DOI: 10.1145/3708531.
- Wu, Y. and Lam, J. S. L. (2026). “Strategic investment in supply chain resilience: A study of fast-moving consumer goods under demand uncertainty”. In: *Transportation Research Part E: Logistics and Transportation Review* 205, p. 104495. DOI: 10.1016/j.tre.2025.104495.
- Xu, Y., Tang, J., and Men A. anf Chen, Q. (2024). “EviPrompt: evidential prompt generation method”. In: *IEEE Transactions on Image Processing*. DOI: 10.1109/TIP.2024.3482175.
- Yeen, H., Wu, R., Juntao, H., and Yingping, X. (2024). “Evaluating ChatGPT-4.0’s data analytic proficiency in epidemiological studies: A comparative analysis with SAS, SPSS, and R”. In: *Journal of Global Health* 14.
- Yerra, S. (2025). “Optimizing Supply Chain Efficiency Using AI-Driven Predictive Analytics in Logistics”. In: *International Journal of Scientific Research in Computer Science Engineering and Information Technology* 11.2, pp. 1212–1220. DOI: 10.32628/cseit25112475.
- Yin, R. (2018). *Case Study Research and Applications: design and methods*. 6th ed. Thousands Oaks, California: SAGE Publications.
- Yu, L., Liga, D., and Markovich, R. (2025). “Addressing the Right to Explanation and the Right to Challenge through Hybrid-AI”. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Law*. ACM, pp. 239–247. DOI: 10.1145/3769126.3769252.
- Yuan, Y., Ding, J., Fen, J., Jin, D., and Li, Y. (2025). “A universal pre-training and prompting framework for general urban spatio-temporal prediction”. In: *IEEE*

Transactions on Knowledge and Data Engineering. DOI: 10.1109/TKDE.2025.3545948.

Zhang, X. and Cavusoglu, H. (July 2025). “Unpacking the Decision Logic of LLM-based Agents: Evidence from the Newsvendor Problem”. In: SSRN Working Paper. DOI: 10.2139/ssrn.5337019. URL: <https://ssrn.com/abstract=5337019>.

Zhao, Y., Xi, W., Fu, X., and Zhao, J. (Aug. 2025). “Enhancing Multimodal Model Robustness Under Missing Modalities via Memory-Driven Prompt Learning”. In: *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*. Ed. by Kwok, James. Main Track. International Joint Conferences on Artificial Intelligence Organization, pp. 2458–2466. DOI: 10.24963/ijcai.2025/274.

Appendix

A Interviewees

Table A.1: Interviewees, their role and system experience

Interviewee	Role	System Experience
H	Application consultant/supply chain consultant	4 years
M	Application specialist	2 years
A	Application consultant	9 years
D	Supply Chain Consultant	15 years
U	Architect	17 years
J	Product owner	20 years
L	Business Architect	3 years

B Interview Questions

The interviews commenced with an explanation of their purpose and scope. It was clarified that the study was limited to the inbound process, and the interviewees were informed about the non-disclosure agreement signed by the authors. The stated motivation for conducting the interviews was to gain a deeper understanding of which aspects should be prioritised and which variables could be incorporated when developing prompts for the AI model. All interviews were conducted in Swedish.

The interviews with the consultants at PipeChain were structured around the following questions:

1. What is your role?
2. What experience do you have with the system
3. What value does your inbound process provide to customers?
4. What aspects should users focus on when using the system?
5. What data is currently available and being measured?
6. Which metrics/KPIs are currently displayed to customers? How are they presented? At what levels are they shown? Which are the most important?
7. What time intervals do customers typically use when comparing KPIs?
8. At what level is the information most important (e.g., supplier, product, delivery)?
9. What would be a reasonable number of "areas" or categories in which deviations should be displayed?

The interviews with the system users at PipeChain's customers were structured around the following questions:

1. Overall, what value do you feel your organization gains from PipeChain's inbound process?
2. What do you typically look at when using the system? What is its primary use case for you?
3. What data is currently available? Which data is used and which is not? How is it used?
4. Which metrics/KPIs are currently displayed to you, and how are they presented?
5. At what levels do you analyze data (eg., per supplier, per product, etc.)? What levels are most important?
6. How do suppliers use the system?
7. What time intervals are used when comparing KPIs (daily, weekly, monthly)?
8. At what levels do you analyze data (eg., per supplier, per product, etc.)?

9. If an AI system were to analyze the data and identify deviations or trends, what would be a reasonable number of alerts or notifications for a user to receive? Do you have any suggestions for what such a notification should look like?
10. Do you see any potential in the system that is not currently being utilized?

C Screening Table

Table C.1: Screening of existing prompting frameworks in relation to PipeChain SCM requirements

Framework	Numerical Data Handling	Explainability (CoT / rea- soning)	Output Structur- ing	Domain Gen- eralizability	Relevance to PC
CPGF (Kan et al., 2023)	Medium	Medium	High	High	Medium
PG-GVLDL (Huang et al., 2026)	Low	Low	Low	Low (domain- specific)	Low
UniST (Yuan et al., 2025)	Medium	Low	Low	Medium	Low
Structured BAS (Lee and Park, 2025)	Low	Low	Medium	Low (visual do- main)	Low
Iterative Stepwise Optimization	Low	High	Medium	Low (medical)	Low
Hybrid-AI Symbolic Constraints	Medium	High	Low	Low (legal)	Low
MemPrompt (Zhao et al., 2025)	Medium	Low	Low	Medium	Low
EviPrompt (Xu et al., 2024)	Low	Low	Low	Low (image segmentation)	Low
Stage-wise Prompt Portfolio (Jiang et al., 2025)	Low	Medium	Medium	Medium	Low
WSPQCF Framework (Jou and Hao, 2026)	Low	Medium	Low	Medium	Low
Agentic Workflow Framework (Kuligin et al., 2025)	Medium	High	Medium	Medium	Medium
SAFE-PE (evaluation framework) (Laeq and Saleem, 2025)	N/A	N/A	N/A	High	High (evalua- tion only)

Note: Relevance is assessed based on the framework’s ability to support numerical data interpretation, deviation detection, explainability, and structured decision-support outputs.

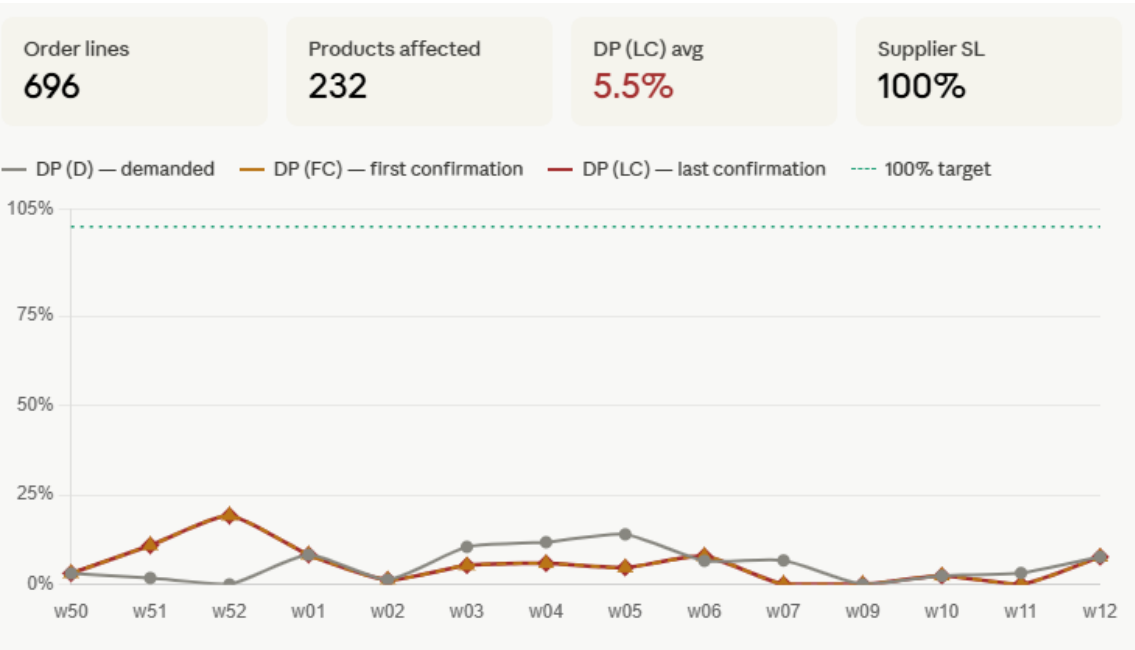
D Output from Theoretical Framework

Output From Theoretical Framework

I have a clear picture of the data. Let me check the chart module and then build the analysis. I have everything I need. The data covers 7,615 order lines across 66 suppliers and 2,021 products from week 50/2025 to week 12/2026. 5,178 lines (68%) are deviations — at least one of DP(D), DP(FC), or DP(LC) is below 100%. Since that’s far more than three, I ranked deviations by a severity score weighted toward DP(LC) (the most crucial KPI), volume of affected lines, and any concurrent service-level dip, then picked the three most severe — all of which are at the supplier level (each spans many products, so the supplier is the right unit of analysis).

Deviation 1 — Supplier 46: chronic late deliveries across the entire portfolio

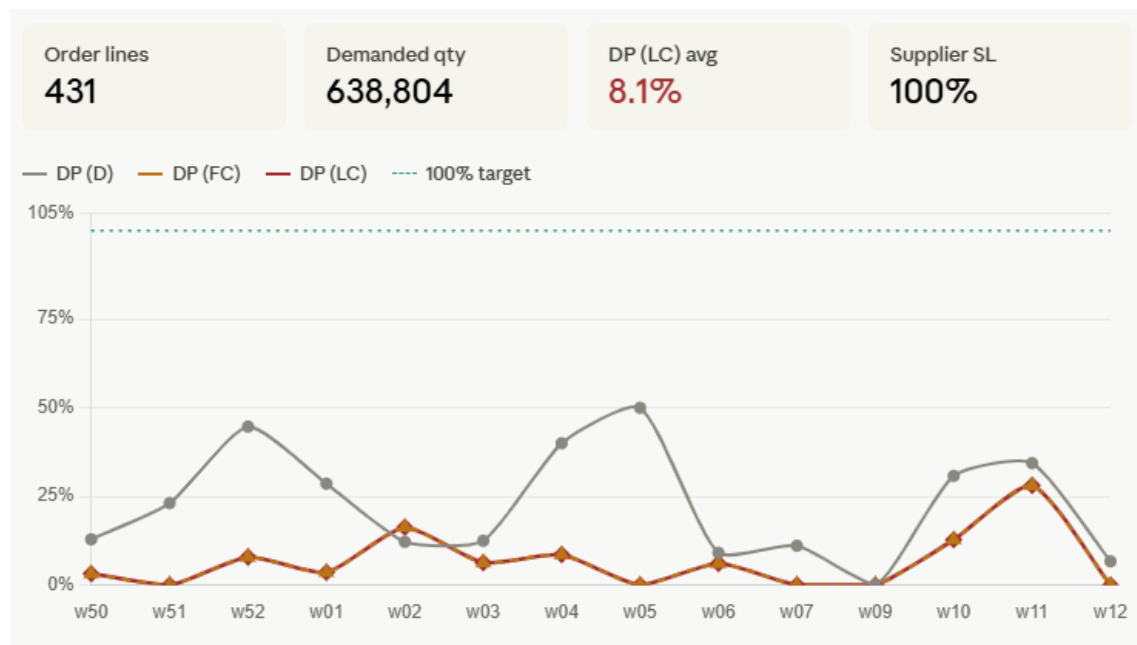
This is the most severe deviation in the dataset. Supplier 46 has 696 order lines spanning 232 products — about 9% of all inbound order volume — and is essentially never on time. DP(D) is 5.0%, DP(FC) is 5.5%, and DP(LC) is 5.5%. The pattern holds week after week with no recovery. Critically, DP(FC) and DP(LC) are identical at the row level, meaning when a confirmation slips, the supplier almost never issues a corrected last confirmation — the customer is left believing the original date until it fails. Of the 232 products this supplier ships, 227 have a DP(LC) below 100%, so the problem is structural, not product-specific. The chart shows the orange (FC) and red (LC) lines tracking almost perfectly on top of each other — this is the visual fingerprint of a supplier that doesn’t re-confirm. The grey D line oscillates a bit because the customer occasionally shifts demand, but FC and LC stay locked together. Quantity-wise, fulfillment is fine (received qty $\approx$ demanded qty), so this isn’t a capacity problem - it’s a timing and communication problem.



Root cause (most likely): A combination of two things. First, the agreed lead time appears systematically shorter than the supplier's actual production cycle, so even the first confirmation is set to a date that can't be met. Second, the supplier's internal process doesn't trigger an updated last confirmation when things slip — the EDI/portal handshake is one-way. Because service level is still 100%, the customer has likely absorbed the lateness by carrying excess safety stock, which masks the financial cost of the deviation. Recommendations: Open a supplier review and renegotiate lead times against demonstrated cycle time; require LC updates whenever an order moves more than X days; check whether safety stock for these 232 products can be safely reduced once the timing stabilizes — that's where the hidden savings sit.

Deviation 2 — Supplier 28: large-volume miss with broken re-confirmation loop

Supplier 28 is the second-largest source of deviation, with 431 lines across 53 products and the largest demanded quantity of any single supplier in the dataset (638,804 units). DP(D) is 23.7%, DP(FC) is 8.1%, and DP(LC) is also 8.1%. The DP(D) & DP(FC) pattern is interesting: it tells you the supplier's first confirmation is often later than the demanded date — they accept the order knowing they can't hit D — and then, like Supplier 46, they fail to update LC when even the slipped FC date is missed. The DP(D) grey line sits well above the FC and LC lines for almost every week — confirming that the supplier is buying slack at the confirmation stage, then still missing the slacker date. Like Supplier 46, FC and LC are identical, so the re-confirmation step is broken.



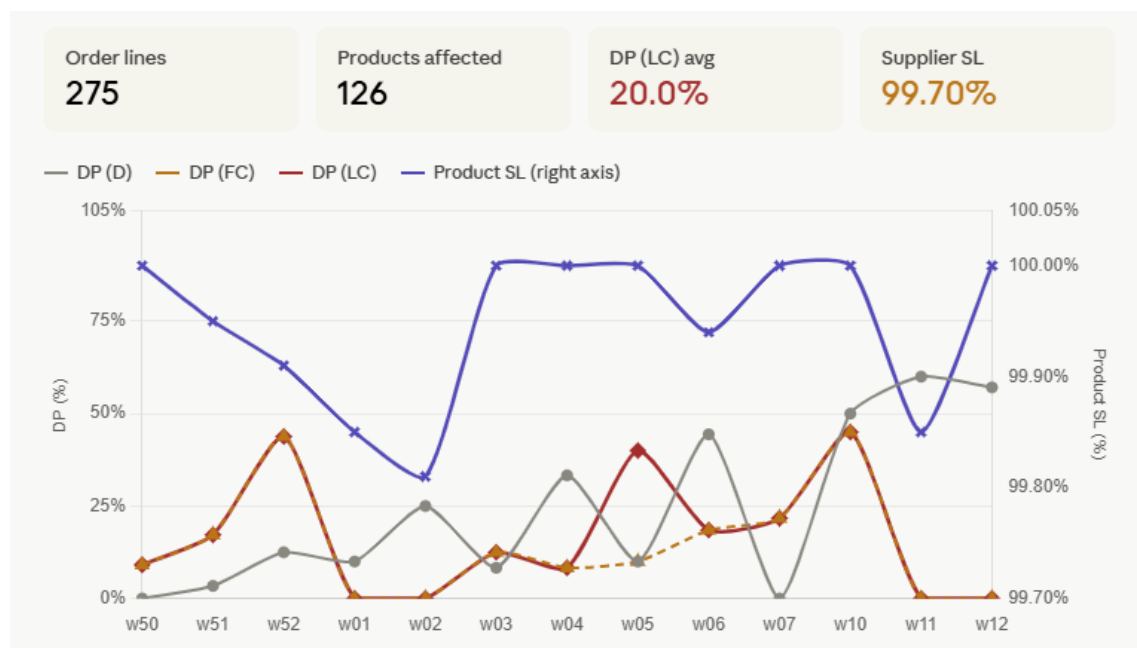
Root cause (most likely): Insufficient capacity or raw-material reliability for this product set, combined with a confirmation process that prioritizes accepting orders over committing to honest dates. The supplier's commercial process and operational

process are out of sync. Service level is still 100% because the customer's safety stock or the supplier's overshipment (received qty is 1.8% above demanded) is buffering the impact.

Recommendations: Treat this as a top-priority supplier escalation given the volume. Request a capacity and lead-time analysis for the 53 products. In parallel, audit the FC vs LC step — if FC and LC are systematically identical, the supplier's portal integration may not be writing LC updates at all, which is a fixable data/process issue separate from the underlying lateness. Track the persistent over-shipment (~12k units) — it could be a sign of batch-size mismatch that's worth addressing commercially.

Deviation 3 — Supplier 23: lateness now visibly degrading service level

Supplier 23 is the deviation where DP problems are no longer being absorbed silently — they're starting to break through to the customer's stock position. 275 lines across 126 products, with DP(D) 26.5%, DP(FC) 18.9%, and DP(LC) 20.0%. The numbers are less extreme than Suppliers 46 and 28, but this is the only top-3 supplier with a Supplier SL below 100% (99.7%) and Product SL dipping at the order-line level. That makes it strategically the most urgent, given that DP-driven stockouts are the user's stated concern. Look at the Product SL line (purple, right axis): it drops below 100% precisely in the cluster of weeks where DP(LC) is at its worst (w51–w02 and again at w11). The dips are small in absolute terms (a few hundredths of a percent) but the directional correlation is the canary you were asked to monitor — DP slippage that the safety stock is no longer fully absorbing. This supplier also shows a worrying trend in the most recent weeks (w10–w12), where DP(D) is climbing into the 50–60% range, meaning the customer is now placing orders that the supplier can hit on the demanded date but the timing on the confirmed dates has gotten worse.



Root cause (most likely): Drift, not collapse. Unlike Suppliers 46 and 28, the FC/LC re-confirmation step does fire here occasionally (e.g. w05 where FC and LC differ), so the data pipeline works. The pattern suggests an operational squeeze — perhaps capacity tightening or input-component shortages on a subset of the 126 products — that’s now eating into the customer’s safety stock buffer. Recommendations: Treat this as the highest-priority operational signal of the three, despite smaller raw numbers, because the SL link means real stockout risk. Drill into the specific products where $DP(LC) = 0\%$ with concurrent SL dips — start with products 159, 236, 268, 266 — and check whether the demand profile changed or the supplier’s lead time has actually shifted. Consider a short-term safety-stock increase on the most exposed products while you investigate, and put Supplier 23 on a weekly DP review until both DP and SL recover.

A few cross-cutting observations worth noting. The $DP(FC) = DP(LC)$ pattern appears for both of the highest-volume suppliers, which strongly suggests a systemic process gap rather than two coincidences — it’s worth checking whether the LC update is even required in the supplier portal’s workflow, because if it’s optional, suppliers will skip it. And the overall 68% deviation rate means this is not a few-bad-apples problem; the three suppliers above are the worst, but Suppliers 22, 41, 10, 19, 42 and 50 also have severe issues and would be the next tier to investigate. If you’d like, I can produce an Excel workbook with the full per-supplier and per-product deviation ranking (so you can sort/filter for ongoing monitoring), or dig deeper into any one supplier’s product mix.

E Output from Adjusted Framework

The following pages contains pictures from the slides provided in the output generated by Clade when conducting T1 in E1.

DEVIATION REPORT

Inbound Order Process · Supply Chain Monitoring

7,615

Order Rows Analysed

5,178

Deviation Rows

45.9%

Overall DP(LC)

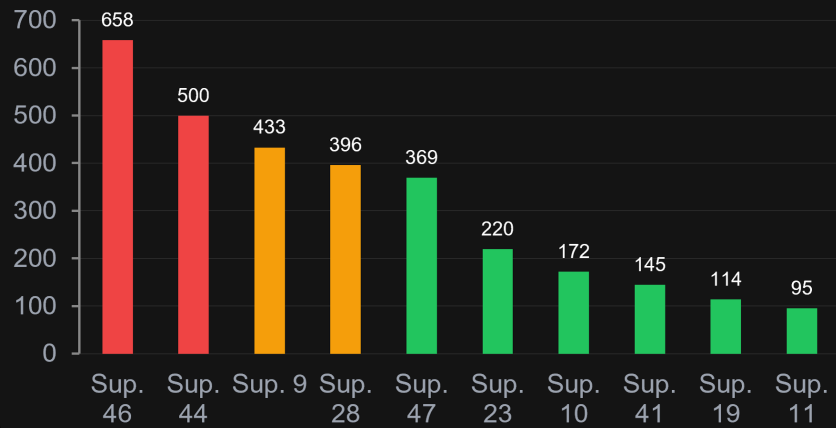
55

Poor-Performing Suppliers

DP(LC) = Last Confirmation Delivery Precision · KPI priority: DP(LC) > DP(FC) > DP(D) · Top 5 suppliers flagged as critical

Supplier Performance Summary

LC Deviations by Supplier (Top 10)



► TOP 5 CRITICAL SUPPLIERS

Sup.	DP(LC)	Criticality
Sup. 46	5.5%	CRITICAL
Sup. 44	50.6%	CRITICAL
Sup. 9	45.2%	HIGH
Sup. 28	8.1%	CRITICAL
Sup. 47	43.1%	HIGH

Sup. 46

Near-total delivery failure across all 14 weeks; DP(LC) = 5.5% on 696 orders.

Sup. 44

Highest order volume (1 012 orders); DP(LC) = 50.6% with deviations spread across 13 weeks and >50 products.

Sup. 9

790 orders, DP(LC) = 45.2%; Supplier SL at 97.4% signals real stock impact. Concentrated in 5 key products.

Sup. 28

DP(LC) = 8.1% but DP(D) = 23.7%, indicating systematic over-confirmation to later dates than demanded.

Sup. 47

648 orders, DP(LC) = 43.1%; consistently poor across all 12 weeks with no improvement trend.

Supplier 46 – Near-Total Delivery Failure

5.0%

DP(D)

5.5%

DP(FC)

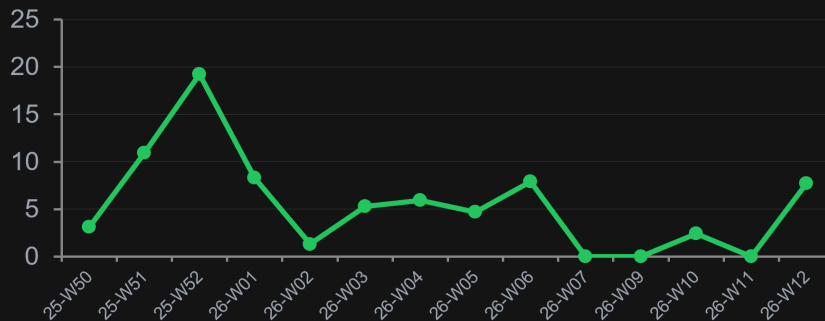
5.5%

DP(LC)

CRITICAL

Criticality

Weekly DP(LC) % — 2025w50 to 2026w12



► TOP PRODUCTS BY DEVIATIONS

Product	Orders	Deviations	DP(LC)
893	16	15	6.3%
1111	15	14	6.7%
686	13	13	0.0%
889	12	12	0.0%
899	11	11	0.0%

► ROOT CAUSE ANALYSIS

DP(D)=DP(FC)=DP(LC) ≈ 5% confirms a systemic, unresolved mismatch between confirmed and actual delivery dates — not a one-time event. The supplier SL remains at 100%, indicating stock levels are maintained, likely through early shipments that arrive outside the confirmed delivery window. This points to a structural lead-time or calendar-alignment issue: orders may be shipped to a fixed internal schedule that does not align with the agreed delivery dates.

► RECOMMENDED ACTIONS

Immediate

Escalate to supplier relationship manager; request urgent explanation for the near-zero DP(LC).

Short-term

Conduct a joint delivery accuracy review; align internal lead times and calendar conventions.

Long-term

Integrate supplier into shared planning system; renegotiate SLAs with automated penalty triggers.

Supplier 44 – Broad Fulfilment Failure Across Product Range

42.3%

DP(D)

50.1%

DP(FC)

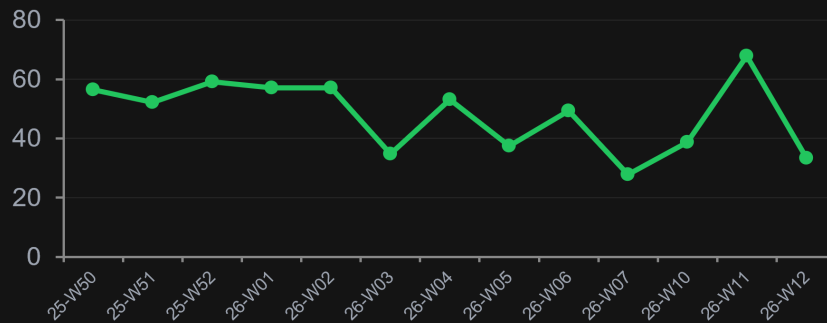
50.6%

DP(LC)

CRITICAL

Criticality

Weekly DP(LC) % — 2025w50 to 2026w12



TOP PRODUCTS BY DEVIATIONS

Product	Orders	Deviations	DP(LC)
581	17	16	5.9%
1612	20	16	20.0%
1041	17	14	17.6%
1511	13	12	7.7%
1652	17	11	35.3%

ROOT CAUSE ANALYSIS

DP(FC) ≈ DP(LC) at ~50%, meaning no course-correction occurs between first and last confirmation. The wide spread across products (no single product dominates) points to a process or capacity constraint affecting the supplier's fulfilment operation broadly rather than a product-specific issue. The consistently moderate DP(LC) (~50%) week on week suggests insufficient planning accuracy or a systematic delay in outbound logistics.

RECOMMENDED ACTIONS

Immediate

Request weekly delivery-status reports until DP(LC) exceeds 80%.

Short-term

Jointly review the supplier's order-to-ship process; identify bottlenecks causing 50% miss rate.

Long-term

Introduce consignment or vendor-managed inventory (VMI) for high-frequency products to reduce lead-time dependency.

Supplier 9 – Recurring Delays Causing Stock Impact

44.6%

DP(D)

45.2%

DP(FC)

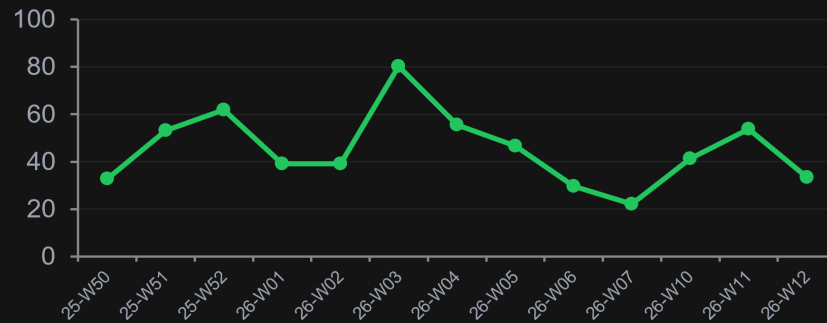
45.2%

DP(LC)

HIGH

Criticality

Weekly DP(LC) % — 2025w50 to 2026w12



► TOP PRODUCTS BY DEVIATIONS

Product	Orders	Deviations	DP(LC)
1047	64	42	34.4%
421	34	23	32.4%
403	39	21	46.2%
1143	34	20	41.2%
369	24	18	25.0%

► ROOT CAUSE ANALYSIS

The slight service level degradation (97.4%) confirms that late deliveries are causing real stock-outs rather than early shipments. Product 1047 alone drives 42 deviations with only 34.4% DP(LC), suggesting capacity constraints or long production cycles for high-runner items. The variable weekly pattern (22% to 80%) indicates reactive rather than planned scheduling on the supplier side.

► RECOMMENDED ACTIONS

Immediate

Flag product 1047 and 421 for safety-stock review; increase buffer inventory.

Short-term

Review capacity utilisation with supplier for top-5 products; set minimum batch planning horizons.

Long-term

Establish rolling 6-week demand signals to supplier; explore dual-sourcing for product 1047.

Supplier 28 – Systemic Confirmation Date Mismatch

23.7%

DP(D)

8.1%

DP(FC)

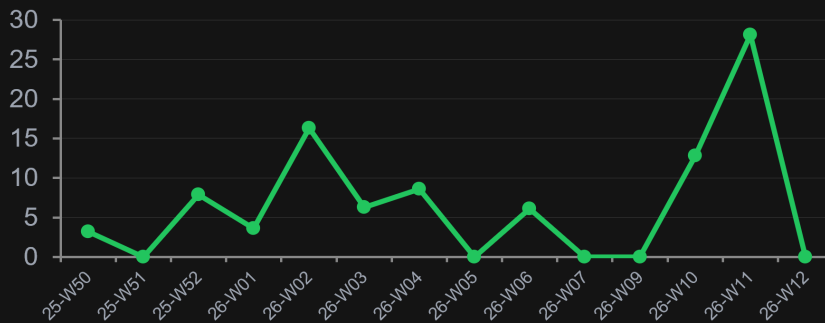
8.1%

DP(LC)

CRITICAL

Criticality

Weekly DP(LC) % — 2025w50 to 2026w12



► TOP PRODUCTS BY DEVIATIONS

Product	Orders	Deviations	DP(LC)
367	50	47	6.0%
366	42	42	0.0%
371	41	38	7.3%
398	31	31	0.0%
982	30	29	3.3%

► ROOT CAUSE ANALYSIS

DP(D) = 23.7% is notably higher than DP(LC) = 8.1%, revealing that the supplier frequently confirms a delivery date later than demanded, then sometimes delivers early (hitting the demanded date) but rarely hitting the confirmed date. This pattern is consistent across all 14 weeks and all top products. Root cause: fixed internal lead times that are longer than agreed lead times, combined with a confirmation process that rubber-stamps the internal schedule rather than the customer's demand date.

► RECOMMENDED ACTIONS

Immediate

Audit the supplier's confirmation process; demand confirmation within 2 business days aligned to DP(D).

Short-term

Renegotiate agreed lead times to reflect actual capability; set DP(LC) KPI target of ≥85%.

Long-term

Integrate ERP order-confirmation visibility; implement automated alerts for LC-vs-demand-date deviations.

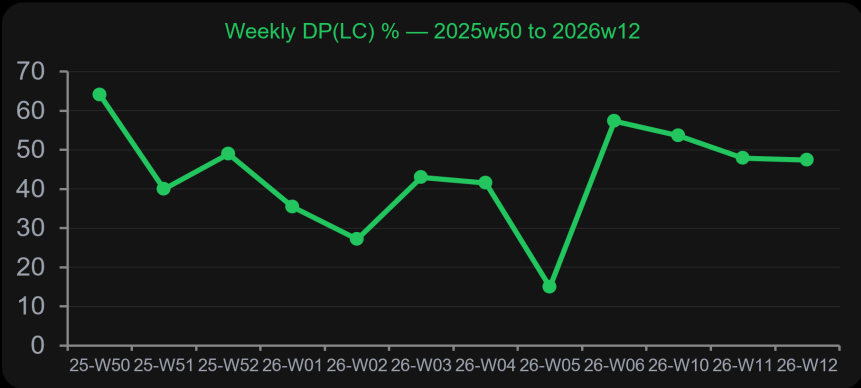
Supplier 47 – Persistent Structural Underperformance

42.1%
DP(D)

43.1%
DP(FC)

43.1%
DP(LC)

HIGH
Criticality



TOP PRODUCTS BY DEVIATIONS

Product	Orders	Deviations	DP(LC)
945	46	28	39.1%
1007	46	28	39.1%
1010	38	23	39.5%
1013	38	21	44.7%
1006	33	20	39.4%

ROOT CAUSE ANALYSIS

DP(FC) ≡ DP(LC) throughout the period — no rescheduling or course-correction occurs after first confirmation. Top-5 products all cluster around 39–44% DP(LC), suggesting the issue is a systemic supply chain or logistics constraint rather than product-specific. The stable ~43% miss rate across 12 weeks and the uniform product impact points to a structural throughput or transport capacity limitation.

RECOMMENDED ACTIONS

- Immediate**

Escalate persistent 43% DP(LC) in monthly review meeting; request corrective action plan.
- Short-term**

Investigate transport and warehousing bottlenecks; explore alternative logistics routes.
- Long-term**

Restructure replenishment cadence for top-5 products; consider safety stock increase of 20–30% pending improvement.

Recommended Actions Summary

IMMEDIATE	SHORT-TERM	LONG-TERM
Supplier 46: Escalate to supplier relationship manager; request urgent explanation for the near-zero DP(LC).	Supplier 46: Conduct a joint delivery accuracy review; align internal lead times and calendar conventions.	Supplier 46: Integrate supplier into shared planning system; renegotiate SLAs with automated penalty triggers.
Supplier 44: Request weekly delivery-status reports until DP(LC) exceeds 80%.	Supplier 44: Jointly review the supplier's order-to-ship process; identify bottlenecks causing 50% miss rate.	Supplier 44: Introduce consignment or vendor-managed inventory (VMI) for high-frequency products to reduce lead-time dependency.
Supplier 9: Flag product 1047 and 421 for safety-stock review; increase buffer inventory.	Supplier 9: Review capacity utilisation with supplier for top-5 products; set minimum batch planning horizons.	Supplier 9: Establish rolling 6-week demand signals to supplier; explore dual-sourcing for product 1047.
Supplier 28: Audit the supplier's confirmation process; demand confirmation within 2 business days aligned to DP(D).	Supplier 28: Renegotiate agreed lead times to reflect actual capability; set DP(LC) KPI target of ≥85%.	Supplier 28: Integrate ERP order-confirmation visibility; implement automated alerts for LC-vs-demand-date deviations.
Supplier 47: Escalate persistent 43% DP(LC) in monthly review meeting; request corrective action plan.	Supplier 47: Investigate transport and warehousing bottlenecks; explore alternative logistics routes.	Supplier 47: Restructure replenishment cadence for top-5 products; consider safety stock increase of 20–30% pending improvement.

F Results from T3

The results from the assessment conducted in T3 are presented in the tables below. Table F.1 and Table F.2 present the results from the assessment of the output generated by ChatGPT. Table F.3 and Table F.4 present the results from the assessment of the output generated by Claude. Table F.5 and Table F.6 present the results from the assessment of the output generated by Gemini. Each table presents the individual scores as well as the average score for each of the four evaluated areas.

Table F.1: The results from the assessments conducted in T3 on the output generated from ChatGPT in E1.

Assessor	Process Interpretation	Root-cause Suggestions	Suggested Actions	Customer Value
1	4	3	2	3
2	5	4	3	4
Average	4,5	3,5	3,5	3,5

Table F.2: The results from the assessments conducted in T3 on the output generated from ChatGPT in E2.

Assessor	Process Interpretation	Root-cause Suggestions	Suggested Actions	Customer Value
1	4	5	4	4
2	3	4	2	4
Average	3,5	4,5	3	4

Table F.3: The results from the assessments conducted in T3 on the output generated from Claude in E1.

Assessor	Process Interpretation	Root-cause Suggestions	Suggested Actions	Customer Value
1	5	5	5	5
2	5	5	4	5
Average	5	5	4,5	5

Table F.4: The results from the assessments conducted in T3 on the output generated from Claude in E2.

Assessor	Process Interpretation	Root-cause Suggestions	Suggested Actions	Customer Value
1	5	5	4	5
2	5	4	3	5
Average	5	4,5	3,5	5

Table F.5: The results from the assessments conducted in T3 on the output generated from Gemini in E1.

Assessor	Process Interpretation	Root-cause Suggestions	Suggested Actions	Customer Value
1	3	2	1	2
2	3	1	1	1
Average	3	1,5	1	1,5

Table F.6: The results from the assessments conducted in T3 on the output generated from Gemini in E2.

Assessor	Process Interpretation	Root-cause Suggestions	Suggested Actions	Customer Value
1	2	1	1	2
2	1	1	1	2
Average	1,5	1	1	2

G Prompts

Prompts for T1, T3, and T4

Below are the prompts used for the calculation accuracy, expert assessments, and consistency of deviation reliability tested in T1, T3, and T4 respectively, presented. First is the prompt for E1 presented, and then the prompt for E2.

Below is the prompt framework used in E1 for T1 and T4:

Table G.1: Prompt framework for T1 and T4 in E1

Purpose	Find deviations in the attached data file, analyse and present them both in text and in a powerpoint.
Context	The data is collected in a supply chain context, where companies and their suppliers share a common database. Customers place an order with a demanded delivery date (D). Suppliers confirm the delivery date two times, first confirmation (FC) and last confirmation (LC). These are compared to D, and delivery precision is computed as a KPI showing how often deliveries arrive at requested dates. DP is measured against D, FC and LC, where LC is the most crucial and D the least. Service level is also available at supplier and product levels. Service level is defined as the fraction of time stock level spends below a crucial threshold. A lacking service level might be a consequence of poor delivery precision. Therefore delivery precision should be prioritized higher when presenting deviations. Deviations are presented to a user who seeks to monitor the purchasing company's inbound order process, and identify which suppliers and products are creating issues.
Input Data Explanation	The data shows placed orders over time, with each row representing one order row. Data can be analyzed on four different levels; supplier, product group, product and order. Product groups and products are sub-groups to suppliers. Time is irrelevant in the column for dmd shipment time.
Task Breakdown	1. Go through each row in the three DP columns. 2. Mark each row that does not consistently have 100% as a deviation. 3. Note the suppliers that are experiencing deviations. 4. Note the poor-performing suppliers and note what products are experiencing deviations. 5. Do a root cause analysis for the deviations. 6. Based on the findings and analysis, give recommended actions for the user.

Continued on next page

Table G.1: Prompt framework for T1 and T4 in E1

Output Specifica- tion	The findings should be presented in a powerpoint: Cosmetics: Black and green colors as theme of the presentation All graphs should have clear axis headers Structure: Slide 1: Front page titled “Deviation Report” presenting: 1. Number of rows analyzed 2. Number of deviation rows 3. Overall DP(LC) percentage 4. Number of poor performing suppliers Slide 2: Summary of supplier performance visualizing suppliers that have the most deviating rows and summarizing the analysis containing: Overall KPI analysis on supplier level Bar chart illustrating the poor performing suppliers. X-axis: supplier Y-axis: number of deviations Highlight the most critical suppliers (maximum 5) listing: 1. DP (LC) 2. Criticality state and short description of why Supplier slides (Number varies based on number of deviations. One slide per supplier): Presentation of each of the deviations, one slide per deviation. The format is: Header Format: “Supplier name - Issue summary” Present why this supplier is identified as problematic through KPI visualization of the available time period in data. If there is a pattern of certain products causing a large part of the supplier’s issues, present those products. Present and motivate possible root cause(s) of the issues Suggest possible actions to manage issues Final slide: Summarize recommended actions for each of the suppliers The actions should be divided into immediate, short-term and long-term, displayed as three different columns In each column actions for each poor performing supplier should be stated as “Supplier X: recommended action in one sentence”
---------------------------------------	---

Below is the prompt framework used in E2 for T1, T3, and T4:

Table G.2: Prompt framework for T1, T3, and T4 in E2

Purpose	Find deviations in the attached data file, analyse and present them both in text and visually.
Context	The data is collected in a supply chain context. The data describes a production scheduling environment, providing data for production orders. Service level is available at product levels. Service level is defined as the fraction of time stock level spends below a crucial threshold. The data also contains metrics for a forecasting process, where forecast accuracy (FA) is compared to actual quantities produced, and is presented as the percentage of forecast error. Similarly, forecast bias (FB) is available, indicating if forecasts tend to be too high or low. Not all products have an established forecasting process. Use the VMI 1-columns to analyze the forecasting metrics. Deviations are presented to a user who seeks to monitor the producing company's production process, and identify which products, and maybe at which plant, are creating issues. Service level is the most crucial metric, which might be caused by a poor forecasting process.
Input Data Explanation	The data shows performance each week over a one year period, with each row representing one product's weekly performance. Data can be analyzed on two different levels; production site and product.
Task Breakdown	1. Go through each row in the SL and FA VMI 1 columns. 2. Mark each row that does not have 100% as a deviation. 3. Note the products that are experiencing deviations. 4. Do a root cause analysis for the deviations. 5. Based on the findings and analysis, give recommended actions for the user.

Continued on next page

Table G.2: Prompt framework for T1, T3, and T4 in E2

Output Specifica- tion	The findings should be presented in a powerpoint: Cosmetics: Black and green colors as theme of the presentation All graphs should have clear axis headers All presented products should be presented using their product ID Structure: Slide 1: Front page titled “Deviation Report” presenting: 1. Overall metric values for the two production sites 2. Number of poor performing products 3. Patterns in found issues Slide 2: Summary of product performance visualizing poor performing products and summarizing the analysis containing: Overall KPI analysis on product level Bar chart illustrating the poor performing products. X-axis: product Y-axis: number of deviations Highlight the most critical products (maximum 5) listing: 1. SL 2. Criticality state and short description of why these products are deemed problematic Issue slides (Number varies based on number of deviations. One slide per two products): Presentation of each of the deviations. The format is: Header Format: “Product name - One sentence summarizing the issue” Present why this product is identified as problematic through KPI visualization of the available time period in data. Present and motivate possible root cause(s) of the issues Suggest possible actions to manage issues Final slide: Summarize recommended actions for each of the products The actions should be immediate In each column actions for each poor performing product should be stated as “Product X: recommended action in one sentence”
---------------------------------------	---

Prompts for T2

Below are the prompts used for the deviation prioritization accuracy tested in T2 presented. As only the task breakdown and output specification were altered, purpose, context, and input data explanation were kept the same as in the prompt used for T1, T3, and T4, these are the only parts included in the tables here. First is the prompt for E1 presented, and then the prompt for E2. The altered parts are written in bold.

Below is the altered parts of the prompt framework used in E1 for T2:

Table G.3: Prompt framework for T2 in E1

Task Breakdown	1. Go through each row in the three DP columns. 2. Mark each row that does not consistently have 100% as a deviation. 3. Note the suppliers that are experiencing deviations. 4. Note the 5 suppliers that have the largest amount of deviating rows and note what products are experiencing deviations. 5. Do a root cause analysis for the deviations. 6. Based on the findings and analysis, give recommended actions for the user.
----------------	---

Continued on next page

Table G.3: Prompt framework for T2 in E1

Output Specification	The findings should be presented in a powerpoint: Cosmetics: Black and green colors as theme of the presentation All graphs should have clear axis headers Structure: Slide 1: Front page titled “Deviation Report” presenting: 1. Number of rows analyzed 2. Number of deviation rows 3. Overall DP(LC) percentage 4. Number of poor performing suppliers Slide 2: Summary of supplier performance visualizing suppliers that have the most deviating rows and summarizing the analysis containing: Overall KPI analysis on supplier level Bar chart illustrating the poor performing suppliers. X-axis: supplier Y-axis: number of deviations Highlight the 5 suppliers with the most deviating rows: 1. DP (LC) 2. Criticality state and short description of why Supplier slides (Number varies based on number of deviations. One slide per supplier): Presentation of each of the deviations, one slide per deviation. The format is: Header Format: “Supplier name - Issue summary” Present why this supplier is identified as problematic through KPI visualization of the available time period in data. If there is a pattern of certain products causing a large part of the supplier’s issues, present those products. Present and motivate possible root cause(s) of the issues. Suggest possible actions to manage issues Final slide: Summarize recommended actions for each of the suppliers The actions should be divided into immediate, short-term and long-term, displayed as three different columns In each column actions for each poor performing supplier should be stated as “Supplier X: recommended action in one sentence”
-----------------------------	--

Below is the altered parts of the prompt framework used in E2 for T2:

Table G.4: Prompt framework for T2 in E2

Task Breakdown	1. Go through each row in the SL and FA VMI 1 columns. 2. Mark each row that does not have 100% as a deviation. 3. Select the top 5 products that experience the most weeks below 100% in service level. 4. Do a root cause analysis for the deviations. 5. Based on the findings and analysis, give recommended actions for the user.
---------------------------	---

Continued on next page

Table G.4: Prompt framework for T2 in E2

Output Specification	The findings should be presented in a powerpoint: Cosmetics: Black and green colors as theme of the presentation All graphs should have clear axis headers All presented products should be presented using their product ID Structure: Slide 1: Front page titled “Deviation Report” presenting: 1. Overall metric values for the two production sites 2. Number of poor performing products 3. Patterns in found issues Slide 2: Summary of product performance visualizing poor performing products and summarizing the analysis containing: Overall KPI analysis on product level Bar chart illustrating the poor performing products. X-axis: product Y-axis: number of deviations Highlight the 5 most critical products listing: 1. SL 2. Criticality state and short description of why these products are deemed problematic Issue slides (Number varies based on number of deviations. One slide per two products): Presentation of each of the deviations. The format is: Header Format: “Product name - One sentence summarizing the issue” Present why this product is identified as problematic through KPI visualization of the available time period in data. Present and motivate possible root cause(s) of the issues Suggest possible actions to manage issues Final slide: Summarize recommended actions for each of the products The actions should be immediate In each column actions for each poor performing product should be stated as “Product X: recommended action in one sentence”
-----------------------------	---

Prompts for T5

Below are the prompts used for the sensitivity testing in T5 presented. As the task breakdown and output specification were not altered, but kept the same as in T1, T3, and T4, these have been excluded from the tables here. First are the three prompt for E1 presented, and then the three prompts for E2. The altered parts are written in bold.

Below is the altered parts of the prompt framework used in E1 for T5, iteration 1:

Table G.5: Adjusted prompt for E1, first iteration of T5

Purpose	Find issues in the attached data file, analyze and present them both in text and visually .
Context	The data is collected in a supply chain context, where data describes an order process between a purchasing company and their suppliers . Customers place an order with a demanded delivery date (D). Suppliers confirm the delivery date two times, first confirmation (FC) and last confirmation (LC). These are compared the requested delivery dates , and delivery precision is computed as a KPI showing how often deliveries arrive at requested dates. DP is measured against D, FC and LC, where LC represents confirmation accuracy and D the accuracy of deliveries arriving on requested dates . Service level is also available at supplier and product levels. Service level is defined as the fraction of time stock level spends below a crucial threshold. A lacking service level might be a cause of poor delivery precision. Therefore delivery precision should be prioritized higher when presenting deviations. Deviations are presented to a user who seeks to monitor the purchasing company's poor performing suppliers, as well as which products seem to be causing issues .
Input Data Explanation	The data shows placed orders over time, with each row representing one order row. Data can be analyzed on two different levels; supplier, product.

Below is the altered parts of the prompt framework used in E1 for T5, iteration 2:

Table G.6: Adjusted prompt for E1, second iteration of T5

Purpose	Identify deviations in the attached dataset , analyze them , and present your findings in both a written summary and a PowerPoint presentation .
Context	The data is collected within a supply chain context, where companies and their suppliers operate within a shared database. Customers place orders with a requested delivery date (D). Suppliers then confirm delivery dates twice: an initial confirmation (FC) and a final confirmation (LC). These dates are compared to the requested delivery date (D) to evaluate delivery precision (DP) , a key performance indicator that measures how often deliveries meet the requested timing . DP is assessed against D, FC, and LC, with the final confirmation (LC) being the most critical reference point and D the least. In addition , service level data is available at both supplier and product levels. Service level is defined as the proportion of time inventory levels fall below a critical threshold. A low service level may be a consequence of poor delivery precision. For this reason , delivery precision should be prioritized when identifying and presenting deviations. The analysis is intended for users who want to monitor the purchasing company’s inbound order process and identify which suppliers and products are causing performance issues.
Input Data Explanation	The dataset consists of historical order data , where each row represents an individual order line. The data can be analyzed across four hierarchical levels: supplier, product group, product, and order. Product groups and products are subcategories within each supplier. It is important to note that time is not a relevant factor in the demanded shipment time column , and should therefore not be considered in the analysis of that field.

Below is the altered parts of the prompt framework used in E1 for T5, iteration 3:

Table G.7: Adjusted prompt for E1, third iteration of T5

Purpose	Detect variations in the data in the attached file, analyze the variations and present them textually and visually in a powerpoint.
Context	The data is gathered in a supply chain context, from a shared database between organizations and their suppliers. Customers request a demanded delivery date (D). Suppliers affirm the delivery date two times, first confirmation (FC) and last confirmation (LC). Delivery precision (DP) is computed as a KPI showing how often FC and LC match D . LC is the most crucial and D the least. Service level data is accessible at supplier and product levels, defined as the fraction of time inventory levels are below a predetermined threshold. Poor delivery precision may lead to low service level . Therefore delivery precision should be of highest priority when presenting deviations. Deviations are to be presented to a user who is aiming to monitor the company’s inbound order process, and identify problematic suppliers and products.
Input Data Explanation	The attached file contains data showing orders over time. Each row represents one order row. Data can be studied on four levels; supplier, product group, product and order. Product groups and products are sub-classes to suppliers. The column dmd shipment time is irrelevant .

Below is the altered parts of the prompt framework used in E2 for T5, iteration 1:

Table G.8: Adjusted prompt for E2, first iteration of T5

Purpose	Analyze the attached dataset to identify deviations and present the findings both in written form and through visualizations .
Context	The data is collected in a supply chain context and reflects a production scheduling environment, containing information on production orders. Service level is available at the product level and is defined as the proportion of time that inventory levels fall below a critical threshold. The dataset also includes forecasting performance metrics. Forecast accuracy (FA) measures the percentage error between forecasted and actual production quantities, while forecast bias (FB) indicates whether forecasts are systematically too high or too low. Not all products have an established forecasting process; therefore, the VMI 1 columns should be used when analyzing forecasting-related metrics. Deviations should be identified with a focus on supporting users who monitor the company’s production process. The goal is to determine which products, and potentially which production sites, are contributing to performance issues. Service level is the most critical metric and should be prioritized in the analysis, as poor forecasting performance may be an underlying cause of low service levels.
Input Data Explanation	The data represents weekly performance over a one-year period, with each row corresponding to a product’s performance in a given week . Analysis can be conducted at two levels: production site and product.

Below is the altered parts of the prompt framework used in E2 for T5, iteration 2:

Table G.9: Adjusted prompt for E2, second iteration of T5

Purpose	Identify inconsistencies in the attached data file. Analyze and present them in text and in a PowerPoint .
Context	The dataset is collected in a supply chain context, describing a production scheduling environment. The data is for production orders. Service level (SL) is defined as the fraction of time stock level spends below a crucial threshold. SL data is available at product levels. There are also metrics for forecasting, where forecast accuracy (FA) is correlated to produced quantities, and calculated as the percentage of forecast error. Forecast bias (FB) is also available, indicating if forecasts tend to be too high or low. Use the VMI 1-columns to analyse the forecasting metrics as not all products have an established forecasting process. The end user is someone who seeks to monitor the producing company’s production process, and identify problematic products, and production plants. Poor forecasting may cause issues with service level. Therefore, SL is the most critical metric.
Input Data Explanation	The dataset contains weekly performance over a one year period. Each row depicts the weekly performance of a product. Data can be analyzed on two levels; production site and product.

Below is the altered parts of the prompt framework used in E2 for T5, iteration 3:

Table G.10: Adjusted prompt for E2, third iteration of T5

Purpose	Detect variations in the data in the attached file, analyze the variations and present them textually and visually in a powerpoint.
Context	The data is generated in a supply chain environment. It describes a production scheduling setting, providing data for placed production orders over time. Service level is available at product levels. Service level is defined as the fraction of time stock level spends below a crucial stock level. The data also contains KPIs for a forecasting process, where produced quantities of a product is compared to forecasted quantities, and is presented as the percentage of forecast error. Similarly, forecast bias (FB) is included, indicating if forecasts tend to be too high or low. Only certain products have a forecasting process integrated. Use the VMI 1-columns to analyze the forecasting metrics. The analysis should be presented to a user who wants to identify poor-performing products in the production planning process, as well as comparing which production plant is causing more issues. Service level is the most crucial metric, which might be caused by a poor forecasting process.
Input Data Explanation	The data shows performance each week over a one year period, with each row displaying one product's weekly performance. Data can be analyzed on two different levels; production site and product.