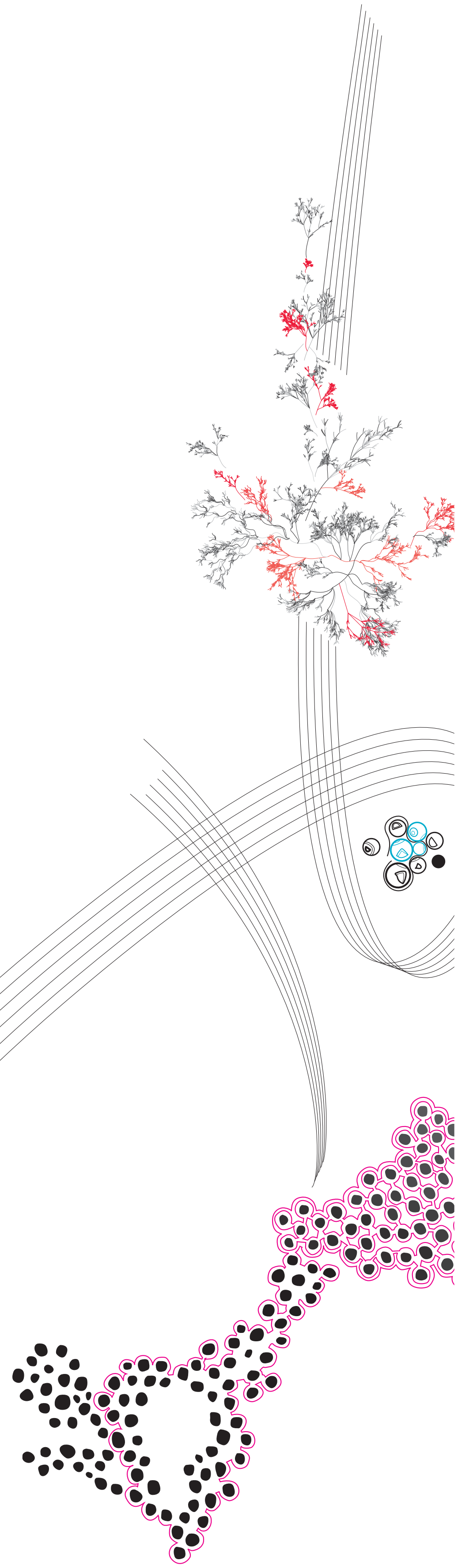

MSc Computer Science
Final Project

Utilizing Generative Artificial Intelligence to enhance Cyber Resilience in Cyber Supply Chain Management

Antónia Zsófia Márton

Supervisor: DR. Abhishta Abhishta

August 28, 2024

Department of Computer Science
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

UNIVERSITY OF TWENTE.

Contents

Abstract

This thesis addresses critical challenges in supply chain management by focusing on the enhancement of cyber resilience and visibility through the integration of N-tier mapping and generative AI. The study identifies the limitations of traditional strategies, which are increasingly inadequate in managing the growing complexity and vulnerability of global supply chains, particularly in the context of evolving cyber threats. The research investigates the transformative potential of Large Language Models (LLMs) when applied within multi-agent systems to automate and optimize supply chain monitoring processes. Although current N-tier mapping frameworks are effective, they can still rely on manual procedures. This thesis demonstrates that the integration of advanced AI technologies can significantly improve the efficiency of these processes.

The findings indicate that automating the N-tier mapping process and leveraging generative AI can substantially enhance supply chain resilience and visibility, leading to more secure and efficient operations. However, the study also underscores the challenges associated with deploying AI in real-world scenarios, highlighting the need for further research and development in this area.

Keywords: cyber security, cyber resilience, cyber supply chain management, generative artificial intelligent

Chapter 1

Introduction

> "The real competition is between supply chains, not companies."
>
> *Martin Christopher*

In a rapid digital transformation era, supply chains — the backbone of the global technological ecosystem — are facing unprecedented vulnerabilities. These vulnerabilities stem from their inherent complexity and deep reliance on information and communications technology (ICT) products and services, as highlighted by Ghadge et al. The current situation is further complicated by the interconnectedness of supply chains. This not only expands the attack surface accessible to global threat actors but also increases the likelihood of cyber incidents at an alarming rate. Such incidents pose a significant risk to businesses and organizations operating in the digital age. It is essential to recognize the importance of securing supply chains to minimize these risks [24].

The "Securing the Supply Chain" report by Accenture delves deeper into the complexities that characterize modern supply networks in the digital age. It notes that as organizations evolve into more agile, digitally-focused, and customer-centered entities, the expansion of external connections and data exchange points dramatically magnifies the scope for potential risks and vulnerabilities. This evolution, while instrumental in driving operational efficiency and customer satisfaction, inadvertently heightens the security risks, underscoring the pressing need for robust cyber resilience measures [2].

Short-term risks include financial, geopolitical, natural disaster, and operational hazards. Long-term risks include concentration and dependency, legal and compliance challenges, and sustainability issues. Together, these problems highlight the urgent need for improved strategies and technologies to help assess, monitor, and reduce risks throughout the supply chain [4].

1.1 Challenges

The field of cyber supply chain management faces many challenges. Crises are increasing in frequency and size, creating serious challenges for supply chains.

1. **Knowledge Gap:** There is a significant knowledge gap among organizational leaders in cybersecurity-related areas, which results in a lack of understanding regarding what aspects require monitoring and how to efficiently combat advanced threats [51].

2. **Complexity of Supply Chains (SCs):** The complexity of contemporary supply

chain systems, compounded by information overload, further complicates the monitoring process. This complexity makes it challenging to identify and respond to potential risks promptly [65].

3. **Poor Visibility:** A concerning 90% of organizations report having poor visibility into their extended supply chains, making it difficult to understand and manage potential risks [1]. Visibility in supply chains is crucial as it enables organizations to track the flow of goods, monitor inventory levels, and oversee transportation activities in real-time. This facilitates improved decision-making and operational performance [59].

4. **Slow Response:** About 80% of organizations take a week or more to evaluate supply chain disruptions, which severely hampers their ability to respond quickly and effectively. Consequently, 54% of executives estimate significant revenue loss due to these disruptions [1].

5. **Reliance on Manual and Semi-Automated Mapping Methods:** Current mapping methods heavily rely on manual processes, which are labor-intensive and prone to errors. Alternatively, when these processes are automated, it often comes at the cost of accuracy, making reliable mapping a significant challenge in maintaining comprehensive supply chain visibility.

6. **Lack of Standardized Evaluation Metrics:** Accurately assessing supply chain visibility is challenging without standardized metrics, leading to difficulties in enhancing visibility levels. Without innovative solutions, organizations may become more susceptible to cyber threats, disruptions, and inefficiencies in their supply chain operations [44].

Addressing these challenges requires innovative approaches that leverage emerging technologies, such as generative AI, to enhance the identification and mitigation of risks, improve visibility, and streamline monitoring. Generative AI offers the potential to revolutionize supply chain cyber resilience by process automation, more robust analysis and providing actionable insights, aligning with the overarching goal of this thesis.

1.2 Proposed Solution

This thesis advocates for the adoption of N-tier mapping alongside the integration of a Generative AI Multi-Agent System (MAS) to confront a multitude of challenges currently plaguing cyber supply chain management. The primary objective is to develop robust strategies that significantly enhance supply chain resilience. N-tier mapping elucidates the intricacies of supplier dependencies across multiple levels, which is pivotal for identifying and monitoring these suppliers to mitigate inherent risks effectively. Concurrently, the Multi-Agent System employs generative AI to automate and optimize the processes of data collection, risk assessment, and compliance monitoring, thus promising a substantial fortification of cyber resilience within supply chains in the digital era.

Additionally, this approach aims to bridge the knowledge gap through the deployment of a Generative AI-powered chatbot. This chatbot is designed to assess the current maturity level of an organization by posing custom questions and providing tailored suggestions. Furthermore, the system enhances the analysis of the value chain, highlighting critical areas and suppliers that require attention and enabling more targeted interventions. It automates these processes using a multi-agent approach, where agents collaborate as a team, thereby enhancing visibility and enabling faster responses to disruptions. The enhanced capabilities of generative AI are adept at managing the complexities of contemporary supply chains.

To address the standardization problem, the proposed N-tier Mapping introduces a structured methodology for visualizing and managing relationships and dependencies across various supplier tiers.

The MAS leverages generative AI technologies to automate the collection, analysis, and processing of data. This automation applies standardized algorithms and processing rules to ensure that data from various sources are consistent and accurately integrated into the decision-making processes.

Generative AI can adapt to evolving standardization norms and automatically update the processing algorithms without manual intervention. This flexibility ensures that the supply chain system remains compliant with the latest standards and best practices.

The combination of N-tier mapping and MAS allows for the seamless integration of data from various sources. This integration is standardized in a way that all data adhere to predefined formats and metrics, facilitating more reliable and meaningful analytics.

To provide insights and analytics based on uniform criteria the MAS can generate standardized reports. These reports are crucial for internal audits, compliance checks, and strategic planning, ensuring that all parts of the organization and external stakeholders are on the same page.

Through these sophisticated methodologies and technologies, the proposed solution not only tackles existing challenges but also paves the way for a more resilient, efficient, and secure cyber supply chain infrastructure.

1.3 Recent Supply Chain Cyber Attacks

A supply chain attack is a type of cyber attack where attackers exploit vulnerabilities in a company's supply chain network, such as suppliers, vendors, or third-party software libraries. This type of attack was behind several high-profile incidents in 2023.

Airbus was compromised in January 2023 through a breached Turkish Airlines employee account. The threat actor, known as USDoD, accessed Airbus's systems and exposed personal data from over 3,000 Airbus vendors, including Rockwell Collins and Thales Group. This data included names, addresses, phone numbers, and email addresses.

The largest refined oil pipeline in the United States, Colonial Pipeline, was attacked in March 2023 through a remote code execution vulnerability in their PulseConnect Secure VPN software. This attack disrupted operations for five days, causing gasoline shortages in the Southeastern U.S. Colonial Pipeline paid a $4.4 million ransom to regain control.

In May 2023, Norton was breached through a zero-day vulnerability in MOVEit Transfer, managed file transfer software used by their parent company, Gen Digital. The attackers gained access to Norton's network and stole employee personal information, threatening to release the data unless a ransom was paid.

UCSF's electronic health record system was compromised in February 2023 due to a vulnerability in Codecov, a code testing software used by Zellis. The breach prevented UCSF clinicians from accessing medical records and scheduling surgeries, leading to cancellations and delays.

Microsoft was compromised in February 2023 due to a vulnerability in Jfrog Artifactory, a binary repository manager. The attackers injected malicious code into Microsoft's software components, allowing them to access networks and steal source code and confidential data [56].

In addition to traditional methods, several advanced AI-powered techniques could be utilized in supply chain cyber attacks. These include the use of deepfake voice technology, generative AI for phishing, AI algorithms for discovering vulnerabilities and evading detection, and AI-driven chatbot phishing scams. These AI-driven techniques highlight the evolving nature of cyber threats, emphasizing the need for enhanced security measures and vigilance in managing supply chain vulnerabilities.

1.4 Research Questions and Objective

How can Generative Artificial Intelligence be utilized to enhance the cyber resilience in supply chain management?

The primary objective of this thesis is to investigate and illustrate how Generative Artificial Intelligence (GEN AI) can be effectively utilized to bolster the cyber resilience of supply chains. In response to evolving cyber threats, this research seeks to identify areas where GEN AI can automate and optimize cybersecurity processes within supply chain management. The overarching goal is to equip stakeholders, including business leaders and cybersecurity professionals, with a comprehensive understanding of the benefits and challenges associated with implementing GEN AI technologies to enhance supply chain security. Focusing on Generative AI for research is motivated by its advanced capabilities.

Firstly, Creative Content Generation is a significant strength of Gen AI, as it can automate the creation of reports, alerts, and scenarios. This capability facilitates the prediction and anticipation of disruptions by providing managers with timely and accurate insights.

Secondly, Gen AI excels in advanced data analysis through its pattern recognition and natural language processing capabilities. By analyzing large and diverse datasets, it can uncover significant trends and anomalies. Leveraging various data sources such as logistics information, supplier communications, and geopolitical developments, it can provide valuable insights to identify potential bottlenecks and emerging threats.

Thirdly, Process Automation and Optimization are streamlined with Gen AI, which can automate repetitive tasks. This enhances efficiency by reducing the workload on supply chain professionals, ensuring consistent and accurate workflows.

Moreover, Gen AI can offer strategic decision support to supply chain managers through smart recommendations. By analyzing market trends, historical data, and supplier performance, it can help in optimizing sourcing strategies and risk mitigation plans, thereby improving overall resilience in supply chain management.

The decision to explore the implementation of Gen AI in enhancing supply chain cyber resilience also stems from the proven historical success of other technologies in strengthening cybersecurity defenses. These established tools have demonstrated their effectiveness in mitigating cyber threats and bolstering risk identification.

Network Traffic Analysis Tools have substantially improved risk and disruption identification by capturing, decoding, and analyzing network data to provide critical insights into potential security threats. By translating data packets into comprehensible formats, these tools enable network administrators to proactively identify and address network anomalies, such as spyware, virus outbreaks, and Denial of Service (DoS) attacks. They enhance the efficiency of intrusion detection by distinguishing malicious patterns in network behavior, helping professionals identify potentially compromised computers and abnormalities that may threaten security standards [60].

Endpoint Detection and Response (EDR) solutions have advanced the identification of risks and disruptions through continuous monitoring of endpoint activities. This ensures timely detection and response to sophisticated cyber threats. By integrating both endpoint and network data, EDR tools provide comprehensive protection against security breaches, allowing swift identification and investigation of potential intrusions. These solutions streamline security investigations by centralizing and automating threat detection and response, thus reducing the need for security engineers to manually analyze data from different systems. Incorporating multiple layers of security, EDR tools offer robust analytics to help security teams detect threats, analyze their potential impact, and initiate automated responses across the organization. This integrated platform also features web threat scanning and external device scanning, offering comprehensive enterprise coverage and uninterrupted defense. By combining prevention, investigation, detection, and response into a single platform, EDR solutions enhance operational efficiency and the organization's security posture. This ensures consistent identification of disruptions and rapid protection against emerging threats [8].

Security Information and Event Management (SIEM) systems have significantly bolstered risk and disruption identification, providing Security Operations Centers (SOCs) with the means for effective real-time incident detection and monitoring. These systems accomplish this by aggregating security events from numerous sources across enterprise networks, normalizing them into a consistent format, and storing them for forensic analysis. By correlating and analyzing these events, SIEM systems empower security teams to detect and identify malicious activities swiftly, offering actionable insights that facilitate rapid incident response. SIEM systems centralize security event data, delivering a comprehensive view of an organization's security posture. This allows the SOC to quickly identify patterns and anomalies that suggest potential security breaches, supporting proactive threat hunting and incident trend analysis to enhance future security measures. Despite operational and technical challenges, their ability to correlate diverse data sources enhances risk identification, enabling organizations to more effectively manage disruptions and maintain robust security defenses [9].

Intrusion Detection and Prevention Systems (IDPS) play a critical role in bolstering risk and disruption identification by effectively detecting and mitigating security threats within computer systems and networks. They are designed to identify potential security threats in real time and implement automated responses, thereby minimizing risks to the systems under surveillance. These systems employ a variety of detection methodologies, including signature-based detection, which identifies known attack patterns, anomaly-based detection that flags unusual network behavior, and stateful protocol analysis to ensure compliance with established protocol norms. In addition, hybrid methods integrate these techniques to create more comprehensive threat detection strategies. By leveraging this combination of methodologies, IDPS can efficiently detect malicious activities that might otherwise remain unnoticed, providing a robust safeguard against disruptions and maintaining network security. The capacity of IDPS to automatically detect and respond to potential threats in real time enhances the organization's ability to respond proactively to emerging security issues, significantly reducing the potential impact of attacks [42].

Machine Learning (ML) and Artificial Intelligence (AI) technologies have played a significant role in enhancing the identification of risks and disruptions, particularly within the Internet of Things (IoT) ecosystem. Their impact on cybersecurity research has empowered algorithms to analyze extensive datasets, revealing patterns that signal potential threats, which in turn has significantly improved the accuracy of threat detection. In particular, machine learning has elevated the effectiveness of advanced IDSs by evaluating traffic patterns, device behavior, and network anomalies to proactively identify threats, even in the intricate and diverse landscape of IoT environments. These technological advancements

underscore the crucial role of AI and ML in strengthening the security infrastructure by swiftly detecting irregularities and preempting emerging threats [7].
To achieve the stated objective, the study will address the following key research questions:

Q1: *What are the current strategies for cyber resilience specifically within cyber supply chain management?*

Q2: *What is the current design and operational framework of N-tier mapping, and what areas could be enhanced to improve cyber resilience?*

Q3: *What innovative approaches can GEN AI offer to enhance the N-tier process?*

Q4: *What are the potential challenges and limitations of integrating GEN AI into this system?*

1.5 Research Design

This study employs a mixed-methods research approach to comprehensively address the research questions concerning the enhancement of cyber resilience in cyber supply chain management through the application of GEN AI. The methodology adopted combines quantitative and qualitative methods to explore the topic.

The methodological foundation of this thesis is rooted in the mixed methods research approach, which combines quantitative and qualitative methodologies. Creswell emphasizes the value of mixed methods in exploring complex research questions. The complexity of cyber resilience and the innovative application of Generative AI in supply chain systems necessitate a research strategy that goes beyond the conventional boundaries of single-method studies. This comprehensive investigation includes a bibliometric analysis and collaborative work with industry professionals, in line with recommendations for mixed-methods research [63, 14].

The bibliometric analysis quantitatively examines the existing literature, highlighting gaps, trends, and best practices related to the topic. This method is pivotal in mapping out the current state of research and establishing a theoretical framework for the study [19].

The analysis followed a systematic, step-by-step process to explore cybersecurity in supply chain management. First, the Scopus database was selected for its comprehensive literature coverage, and VOSviewer was utilized to visualize and analyze the bibliometric data. Carefully developed search queries were created with various keyword combinations to widen the research scope related to supply chain cybersecurity. Search results were then refined through filters for language (English), fields of study (Computer Science and Engineering), and publication years (2018-2024) to ensure the data's relevance and timeliness.

It was followed by tracking publication trends over time to map the research's growth trajectory and conducting citation impact assessments to identify influential works and their networks using citation counts and Total Link Strength (TLS). This highlighted key contributors who have shaped the field. Co-citation analysis further revealed related research clusters, illuminating the intellectual structure and interrelationships among scholarly papers.

Additionally, bibliographic coupling was conducted to identify documents sharing common references, showcasing the interconnectedness of studies based on shared scholarly foundations. A co-occurrence analysis of author-provided keywords discerned prevalent themes, while a comprehensive keyword analysis identified broader trends.

These systematic steps offered a thorough examination of the academic discourse surrounding supply chain cybersecurity, forming a strong foundation for further research.

On the other hand, collaboration with industry professionals provides practical insights and expert perspectives to enrich the research. This collaboration enhances the overall analysis, offering a deeper understanding of how GEN AI can enhance cyber resilience. Additionally, GEN AI experiments will be conducted toward the end of the thesis to evaluate different inputs, existing models, and prompts. This iterative process aims to deepen our understanding of how GEN AI can be implemented to enhance cyber resilience in supply chain monitoring systems.

1.6 Structure

This thesis is structured into five main chapters, each designed to build upon the insights of the previous one to deepen understanding of cyber resilience in supply chains using generative AI and N-tier mapping. The Introduction sets the context, outlines the significant need for enhanced cybersecurity, and presents the research questions. Chapter 2 offers a literature review with a detailed bibliometric analysis, assessing current knowledge and identifying gaps in the field. In Chapter 3, we explore the application of N-tier mapping and LLMs to improve visibility and responsiveness. Chapter 4 details the experimental methods and outcomes of practical AI applications in supply chain scenarios. Finally, the Discussion and Conclusion (Chapters 5 and 6) evaluate the challenges and limitations, and broader implications of the findings, proposing directions for future research.

Chapter 2

Review of the Literature

2.1 Bibliometric Analysis

Bibliometric analysis has established itself as a robust tool for navigating the intricate field of supply chain management. This methodology offers a structured approach to evaluating the extensive corpus of scientific literature, enabling the identification of significant trends, patterns, research gaps, and the most relevant studies for consideration.

Bibliometric analysis is a methodological approach designed to handle and analyze large volumes of scientific data. This technique facilitates the mapping of the intellectual structure and the evolutionary subtleties of specific domains, thereby illuminating emerging areas of research. The distinct advantage of bibliometric analysis lies in its ability to systematically aggregate, analyze, and present scientific data, facilitating a comprehensive overview of the subject matter at hand [19]. It plays a crucial role in revealing the thematic focuses within the literature, tracing the progression of research themes over time, and delineating the connections across diverse research areas [55]. Furthermore, bibliometric tools such as VOSviewer and Scopus offer advanced data visualization and network analysis capabilities, enhancing the depth and clarity of insights derived from the analysis.

Within the scope of this thesis, the employment of bibliometric analysis through tools like Scopus and VOSviewer is a pivotal step toward developing an in-depth overview of the existing research on supply chain risk management and cyber resilience. By methodically charting the scientific contributions in this field, the analysis seeks to pinpoint key research trends, highlight influential studies, and identify potential gaps in the literature. This structured exploration is essential for laying the groundwork for subsequent research and directing the investigation toward areas poised for substantial academic and practical contributions.

2.1.1 AI's Role in Supply Chain Cyber Resilience

The analysis of queries presented in Table 2.1 provides significant insights into the current literature landscape at the intersection of supply chain management and cybersecurity. A thorough examination of the initial queries (RQ1, RQ2, RQ3, and RQ4) reveals notable variations in result counts, attributable to differences in keyword phrasing.

The progression from RQ1 to RQ2 in Table 2.1, where the keywords extend beyond basic supply chain security to include "threat detection" demonstrates how minor modifications in terminology can significantly expand the research scope. The result count increases from 463 to 1530, indicating a marked escalation in literature focusing on threat detection within supply chain cybersecurity. This surge underscores the increasing scholarly focus on “threat detection” highlighting the critical need for proactive strategies to secure supply chain operations.

The introduction of “industrial” as a keyword in Table 2.1 RQ3 further escalates the result count to 3157. This significant increase reveals the extensive research interest in industrial supply chain systems’ security aspects, suggesting a robust body of work dedicated to addressing cybersecurity challenges in an industrial context.

The inclusion of Information and Communications Technology (ICT) and supply chain risk management terms in Table 2.1. RQ4 query marks a strategic expansion of our research scope, aiming to gain insight into the supply chain management trends. This modest increase suggests that the field of ICT supply chain risk management is indeed relevant but not as extensively explored as the practical, operational aspects of supply chain security. This observation aligns with our research objective to investigate the potential of GEN AI in enhancing cyber resilience within supply chains, indicating a niche where GEN AI could potentially bridge existing gaps in cybersecurity management strategies. By integrating these terms, the aim is to understand the full spectrum of cybersecurity challenges facing supply chains today, thereby identifying where GEN AI can introduce innovative solutions to improve resilience and risk management in a digitalizing global supply chain landscape.

The inclusion of Generative AI-related terms in Q5 marks a crucial shift, specifically focusing on the application of GEN AI technologies in supply chain cybersecurity. The notably lower result count (3) compared to earlier queries highlights that the exploration of GEN AI applications in enhancing cyber resilience in supply chains is still in its infancy, with limited literature available on the subject. This scarcity points to a significant opportunity for innovative research and practical applications in this emerging field.

Table 2.1: Bibliometric Analysis Query Table

ID	Keywords/Phrases	Query Structure	Result Count
RQ1	supply chain, monitoring, visibility, cyber, information, security, cyber resilience	TITLE-ABS-KEY ((“supply chain” AND (monitoring OR visibility)) AND ((cyber OR information) AND security OR “cyber resilience”))	463
RQ2	supply chain, monitoring, visibility, threat detection, cyber, information, security, cyber resilience	TITLE-ABS-KEY ((“supply chain” AND (monitoring OR visibility)) OR “threat detection” AND ((cyber OR information) AND security OR “cyber resilience”))	1543
RQ3	supply chain, industrial, monitoring, supply chain visibility, cyber, information, security, cyber resilience	TITLE-ABS-KEY (((“supply chain” OR industrial) AND monitoring) OR “supply chain visibility” OR “threat detection” AND ((cyber OR information) AND security OR “cyber resilience”))	3157

Table 2.1 continued from previous page

ID	Keywords/Phrases	Query Structure	Result Count
RQ4	ICT, supply chain, risk management, ICT-SCRM, C-SCRM, CSCRM, SCCRM industrial, monitoring, supply chain visibility, threat detection, cyber, information, security, cyber resilience	TITLE-ABS-KEY ((((ict OR cyber) AND "supply chain") AND risk AND management) OR "ICT-SCRM" OR "C-SCRM" OR "CSCRM" OR "SC-CRM" OR (((("supply chain" OR industrial) AND monitoring) OR "supply chain visibility" OR ("threat detection")) AND ((cyber OR information) AND security OR "cyber resilience")))	3483
RQ5	ICT, supply chain, risk management, ICT-SCRM, C-SCRM, CSCRM, SCCRM industrial, monitoring, supply chain visibility, threat detection, cyber, information, security, cyber resilience, Generative AI, GPT, GEN AI, artificial intelligence, AI-driven, AI-based	TITLE-ABS-KEY ((((ict OR cyber) AND "supply chain") AND risk AND management) OR "ICT-SCRM" OR "C-SCRM" OR "CSCRM" OR "SC-CRM" OR (((("supply chain" OR industrial) AND monitoring) OR "supply chain visibility" OR ("threat detection")) AND ((cyber OR information) AND security OR "cyber resilience")) AND (((generative OR gen) AND (ai OR "artificial intelligence")) OR gpt))	3

The analysis of queries reveals an increasingly rich research environment at the intersection of supply chain management and cybersecurity. As illustrated in Table 2.1, there is a foundational body of literature, however, the segment directly addressing the integration of Generative AI to enhance cyber resilience in supply chains remains remarkably underdeveloped. The limited focus on the role of Generative AI in strengthening the cyber resilience of supply chain monitoring systems suggests significant potential for groundbreaking research and practical advancements.

2.1.2 Year on year publication analysis

Before delving deeper into the analysis, this section offers a visual representation of the publication trends over time. Figure 2.1 displays the volume of literature on a year-by-year basis, from the list obtained from Table 2.1 RQ4 query, which does not include terms related to GEN AI.

The graph illustrates a modest increase in publications from the early 1990s until around 2007, suggesting a foundational period of research activity. This period of sustained expansion likely reflects the initial acknowledgement of cybersecurity threats to supply chain systems, and the early development of security measures.

A notable inflection in publication rates emerges around 2007, where the volume begins to escalate more markedly. It might reflect the increasing digitization of supply chains and the resulting new vulnerabilities, drawing greater academic and industry focus to this field.

The period following 2015 is characterized by an exponential increase in the number of publications, as depicted by the sharp uptick in Figure 2.1 This exponential growth phase indicates a burgeoning interest in supply chain cybersecurity, likely driven by the

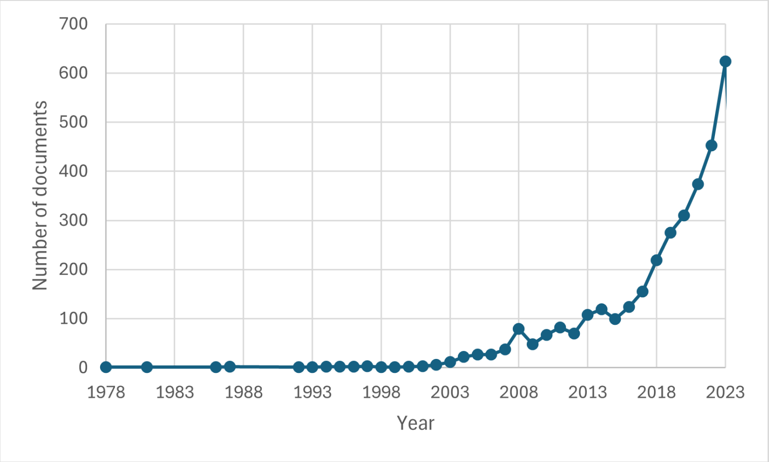

Figure 2.1: Overview of publications per year. Own illustration.

globalization of supply networks, the expansion of digital supply chain solutions, and the evolving sophistication of cyber threats.

By 2023, the peak in publication volume suggests that supply chain cybersecurity has become a field of critical and expanding interest, reflecting a broader recognition of the strategic importance of supply chains to global commerce and national security. The intensified research output may also result from the realization that traditional cybersecurity methods are no longer adequate in the face of advanced persistent threats, necessitating more comprehensive and forward-looking approaches. The data indicates a continued expansion in the field, thanks to the escalating prevalence and impact of cyber threats within the global supply chain landscape.

To conduct the following analysis, the list generated by running the RQ4 query on Table 2.1 was narrowed down to documents related to the subject areas of Computer Science and Engineering in English, published between 2018 and 2024. 2078 documents met this criteria.

2.1.3 Citation analysis based on documents

This subsection showcases the citations-based networks among published documents from the past 6 years. The citations of these documents were analyzed using a threshold of 10 citations per document. Only 447 from 2078 publications were found to satisfy this criterion.

Table 2.2 lists the top ten documents with the highest citation counts and outlines the number of citation-based links each has with other articles within the supply chain cybersecurity niche that also meet the minimum citation criterion. The links between documents in this network are indicative of co-citation patterns identified by VOSviewer. It should be noted that while the links in this citation network suggest patterns of citation by other documents, a comprehensive co-citation analysis will follow later in this document. For example, “The impact of digital technology and Industry 4.0 on the ripple effect and supply chain risk analytics” [18]. received a total of 960 citations, yet only 8 of these are connected to other supply chain cybersecurity articles in Scopus that have ten or more citations.

Figure 2.2 depicts the normalized citation network of the supply chain cybersecurity publications. The network showcases key works of Table 2.2, such as those by Ivanov et al. [18], Ding et al. [17], [16], Muhammad et al. [43], and Mondal et al. [41], as major nodes, each highlighted in distinct colors to represent their unique positions and connections in the research landscape. Notably, the works of Ivanov et al. [18], Ding et al. [17], [16], Muhammad et al. [43], and Mondal et al. [41] are prominent nodes within the network, represented in red, dark blue, orange, and yellow respectively.

Simultaneously, the normalization process has highlighted emerging contributions that, despite their more recent entry into the academic discourse, command significant attention within their clusters. Such works, including those by Lan et al. [36] and Zhou et al. [68], represent fresh and influential perspectives, especially in the realms of tracking technology and industrial systems optimization. Other newly accentuated works, such as those by Haghighi et al. [30] on intrusion prevention and Mihai et al. [40] on digital twins, articulate the technological forefront of the cybersecurity dialogue. The node sizes, within their respective clusters, illuminate these documents' growing centrality to the discussion, with the potential to define new directions for research within the cybersecurity landscape.

Table 2.2: Top 10 cited document with the maximum citation values in SC cybersecurity documents

Rank	**Title of Document**	**Authors**	**Citations**	**Links**
1	The impact of digital technology and Industry 4.0 on the ripple effect and supply chain risk analytics [18]	Ivanov, D., Dolgui, A., Sokolov, B.	960	8
2	A survey on security control and attack detection for industrial cyber-physical systems [17]	Ding, D., Han, Q.-L., Xiang, Y., Ge, X., Zhang, X.-M.	702	7
3	Applying blockchain technology to improve agri-food traceability: A review of development methods, benefits and challenges [22]	Feng, H., Wang, X., Duan, Y., Zhang, J., Zhang, X.	438	0
4	A Survey on Model-Based Distributed Control and Filtering for Industrial Cyber-Physical Systems [16]	Ding, D., Han, Q.-L., Wang, Z., Ge, X.	383	2
5	A Survey of Physics-Based Attack Detection in Cyber-Physical Systems [25]	Giraldo, J., Urbina, D., Cardenas, A., Sandberg, H., Candell, R.	288	0
6	The role of Information and Communication Technologies in Healthcare: taxonomies, perspectives, and challenges [5]	Aceto, G., Persico, V., Pescapé, A.	252	0
7	Secure Surveillance Framework for IoT Systems Using Probabilistic Image Encryption [43]	Muhammad, K., Hamza, R., Ahmad, J., Wang, H., Baik, S.W.	244	2
8	Applications of Wireless Sensor Networks and Internet of Things Frameworks in the Industry Revolution 4.0: A Systematic Literature Review [37]	Majid, M., Habib, S., Javed, A.R., Gadekallu, T.R., Lin, J.C.-W.	235	0
9	Anatomy of Threats to the Internet of Things [38]	Makhdoom, I., Abolhasan, M., Lipman, J., Liu, R.P., Ni, W.	234	0

Table 2.2 continued from previous page

Rank	Title of Document	Authors	Citations	Links
10	Blockchain Inspired RFID-Based Information Architecture for Food Supply Chain [41]	Mondal, S., Wijewardena, K.P., Karuppuswami, S., Kumar, D., Chahal, P.	231	4

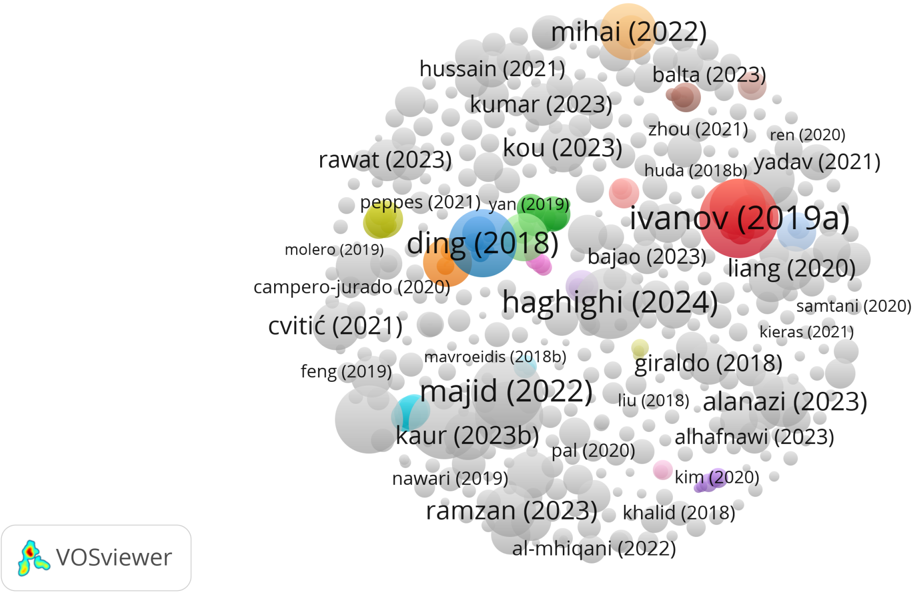

Figure 2.2: Citation network in the field of SC cybersecurity (VOSviewer)

Co-citation

This subsection presents co-citation analysis of publications over the past six years, specifically curated from the Scopus database to uncover the interconnections within the field of supply chain cybersecurity. The initial VOSviewer search culminated in a corpus of 66403 cited research documents, from which a fractional counting method was applied. The methodological threshold was meticulously set to a minimum of five citations, narrowing the dataset to 105 prominently cited documents that form the intellectual backbone of this analysis.

Table 2.3: Top 10 co-cited document with the maximum TLS values in SC cybersecurity documents

Rank	Title of Document	Authors	Year	Citations	TLS
1	Deep learning for unsupervised insider threat detection in structured cybersecurity data streams [61]	Tuor, A., Kaplan, S., Hutchinson, B., Nichols, N., Robinson, S.	2017	14	7

Table 2.3 continued from previous page

Rank	Title of Document	Authors	Year	Citations	TLS
2	Lessons from Stuxnet [11]	Chen, Thomas M., Abu-Nimeh, Saeed	2011	8	7
3	Blockchains and smart contracts for the internet of things [13]	Christidis K., Devetsikiotis M.,	2016	12	6
4	Bridging the gap: a pragmatic approach to generating insider threat data [27]	Glasser J., Lindauer B.,	2013	9	6
5	Insider threat detection with deep neural network [67]	Yuan F., Cao Y., Shang Y., Liu Y., Tan J., Fang B.,	2018	7	6
6	The impact of digital technology and industry 4.0 on the ripple effect and supply chain risk analytics [18]	Ivanov D., Dolgui A., Sokolov B.	2019	6	6
7	A dynamic model and an algorithm for short-term supply chain scheduling in the smart factory industry 4.0 [34]	Ivanov D., Sokolov B., Dolgui A., Werner F., Ivanova M.	2016	6	6
8	Deeplog: Anomaly detection and diagnosis from system logs through deep learning [20]	Du M., Li F., Zheng G., Srikumar V.	2017	8	5
9	XGBoost: A scalable tree boosting system [12]	Chen, T., Guestrin, C.	2016	7	5
10	Behavioral based insider threat detection using deep learning [45]	Nasir R., Afzal M., Latif R., Iqbal W.	2021	7	5

Table 2.3 presents the top ten documents from this group, selected based on the highest Total Link Strength (TLS) values. This metric measures the strength of a document's connections to others within the same citing articles. If 'n' documents are co-cited by a single article, each pair's link strength is calculated as $1/n$. The TLS for each document is the sum of these strengths across all articles where it is cited [55].

Notably, the TLS values differ from the total citation counts. This is because the TLS only accounts for co-citations within our dataset. For example, the paper by Tuor et al. [61] has a TLS of 7, despite being cited 14 times. This implies that there are 7 articles that cite this document but do not cite others from our selected dataset or cite papers with fewer than five citations, thus not meeting our threshold for analysis.

Figure 2.3 of the co-citation network offers a clear depiction of how these documents are interrelated. Tuor et al. [61], Glasser et al. [27], and Yuan et al. [67] converge within a light blue-hued network. The work of Chen et al. [11] emerges as a central node within the brown cluster, while the contributions of Ivanov et al. [18], [34] form the core of an orange-colored network. The other top-cited documents manifest their connections within networks in dark blue, purple, and green.

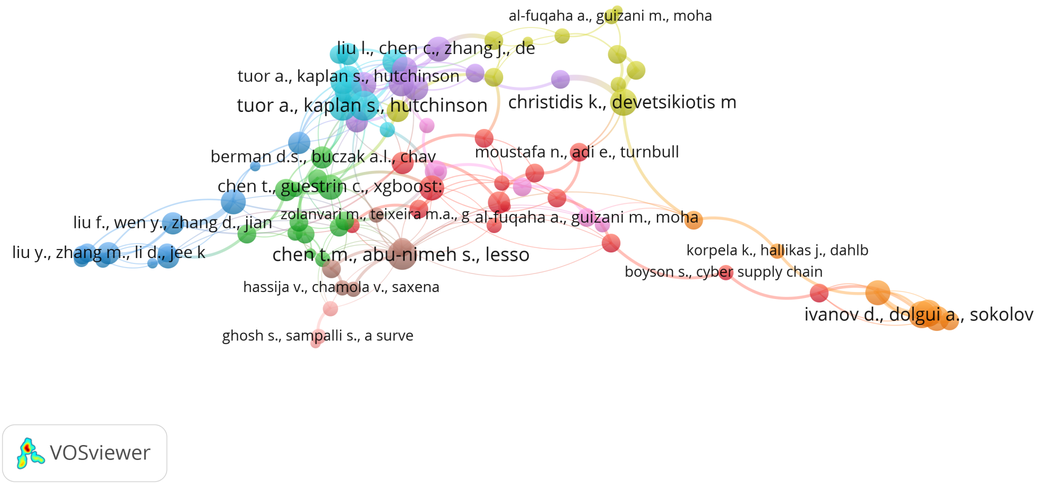
al-fuqaha a., guizani m., moha
liu l., chen c., zhang j., de
tuor a., kaplan s., hutchinson
tuor a., kaplan s., hutchinson
christidis k., devetsikiotis m
berman d.s., buczak a.l., chav
moustafa n., adi e., turnbull
chen t., guestrin c., xgboost:
liu f., wen y., zhang d., jian
zolanvari m., teixeira m.a., g
al-fuqaha a., guizani m., moha
liu y., zhang m., li d., jee k
chen t.m., abu-nimeh s., lesso
korpela k., hallikas j., dahlb
boyson s., cyber supply chain
hassija v., chamola v., saxena
ivanov d., dolgui a., sokolov
ghosh s., sampalli s., a surve
VOSviewer

Figure 2.3: Co-citation network in the field of SC cybersecurity (VOSviewer)

2.1.4 Bibliographic coupling

This subsection explores bibliographic coupling among publications focused on supply chain cybersecurity, drawing from a curated set of 2078 documents indexed in the Scopus database over the past six years. Employing a fractional counting method, the analysis identifies strong connections between documents by setting a methodological threshold of a minimum of 10 citations, narrowing the focus to 447 cited documents.

Table 2.4: Top 10 document with the maximum TLS values in SC cybersecurity documents

Rank	Title of Document	Authors	Year	Citations	TLS
1	Digital Supply Chain Twins: Managing the Ripple Effect, Resilience, and Disruption Risks by Data-Driven Optimization, Simulation, and Visibility [33]	Ivanov, D., Dolgui, A., Das, A., Sokolov, B.	2019	127	59.5
2	The impact of digital technology and Industry 4.0 on the ripple effect and supply chain risk analytics [18]	Ivanov D., Dolgui A., Sokolov B.	2019	960	55
3	An ISM Modeling of Barriers for Blockchain/Distributed Ledger Technology Adoption in Supply Chains towards Cybersecurity [21]	Etemadi, N., Van Gelder, P., Strozzi, F.	2021	39	49
4	A Survey on Digital Twin for Industrial Internet of Things: Applications, Technologies and Tools [66]	Xu, H., Wu, J., Pan, Q., Guan, X., Guizani, M.	2023	12	47

Table 2.4 continued from previous page

Rank	Title of Document	Authors	Year	Citations	TLS
5	Improving supply chain resilience through industry 4.0: A systematic literature review under the impressions of the COVID-19 pandemic [58]	Spieske, A., Birkel, H.	2021	199	41
6	A Survey of Physics-Based Attack Detection in Cyber-Physical Systems [26]	Giraldo, J., Urbina, D., Cardenas, A., Sandberg, H., Candell, R.	2018	288	40
7	A Review of Insider Threat Detection Approaches With IoT Perspective [35]	Kim, A., Oh, J., Ryu, J., Lee, K.	2020	44	38.67
8	A survey on security control and attack detection for industrial cyber-physical systems [17]	Ding, D., Han, Q.-L., Xiang, Y., Ge, X., Zhang, X.-M.	2018	702	38
9	Cyber risk at the edge: current and future trends on cyber risk analytics and artificial intelligence in the industrial internet of things and industry 4.0 supply chains [53]	Radanliev, P., De Roure, D., Page, K., Maddox, L.T., Burnap, P.	2020	63	37
10	Managing Disruptions and the Ripple Effect in Digital Supply Chains: Empirical Case Studies [15]	Das, A., Gottlieb, S., Ivanov, D.	2019	26	35

Table 2.4 highlights the top ten documents ranked by Total Link Strength (TLS). TLS quantifies the cumulative strength of bibliographic connections between documents, based on their shared references. In the fractional counting method, the strength of each link (each shared reference) is divided by the total number of references in the citing document, ensuring each reference contributes equally, regardless of the overall number of references.

For instance, if two documents, A and B, cite a third document C, and A contains 100 references while B contains 10, the link strength from A to C would be 0.01 (1/100), and from B to C would be 0.1 (1/10). The TLS value is then the sum of these fractional values for all documents that a pair of documents co-cites. This metric provides a normalized measure of the intellectual connection between documents, indicating a robust bibliographic linkage when the TLS is high.

The TLS in this context is the sum of these fractional values for all the documents that a pair of documents co-cites. It provides a normalized indication of how strongly two documents are related in terms of their shared intellectual base. The higher the TLS, the more robust their bibliographic connection is, implying that they draw on similar foundational research.

Notably, the documents by Ivanov et al. have the highest TLS values. Interestingly, the document titled “Digital Supply Chain Twins: Managing the Ripple Effect, Resilience, and Disruption Risks by Data-Driven Optimization, Simulation, and Visibility” [33] has a TLS of 59.5 and 127 citations, compared to the document titled “The impact of digital technology and Industry 4.0 on the ripple effect and supply chain risk analytics” [18], which has a lower TLS of 55 but a significantly higher citation count of 960. This discrepancy highlights the varying degrees of bibliographic coupling strength relative to the citation impact. The document by Xu et al. also stands out with a TLS of 47, despite having only 12 citations, highlighting significant bibliographic connections despite fewer citations [66].

Figure 2.4 displays the bibliographic coupling network of the largest connected dataset, which includes 384 documents. The network prominently features documents by Ivanov et al., notably highlighted within an orange cluster. Additionally, key works by Spieske et al. [58], Radanliev et al. [53], and Das et al. [15] also form integral parts of this core orange cluster. Furthermore, documents by Etemadi et al. [21], Xu et al. [66], Giraldo et al. [26], Kim et al. [35], and Ding et al. [17] establish central nodes in the network, each distinctly color-coded in purple, light orange, red, light green, and light purple, respectively. This configuration visually represents the strong bibliographic ties and shared research foundations within the field of supply chain cybersecurity.

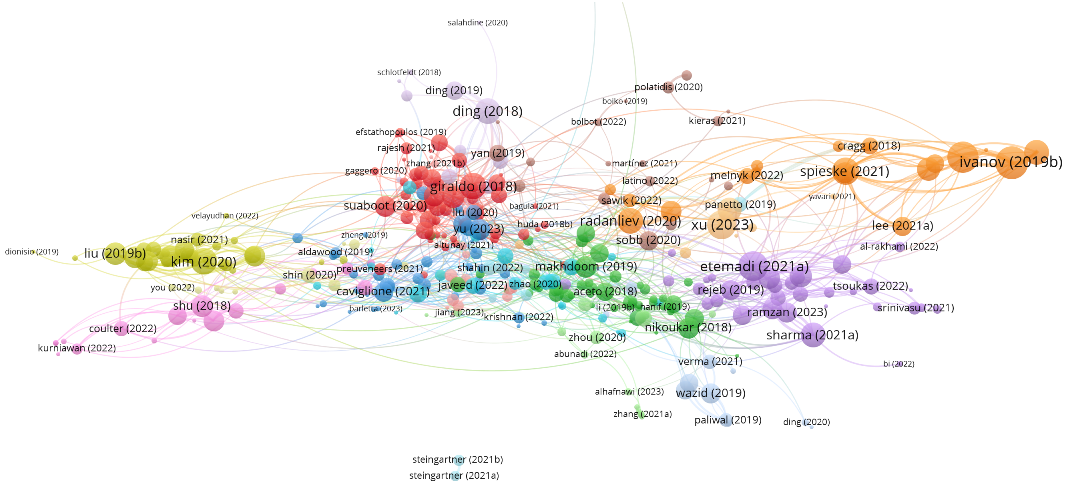

Figure 2.4: Bibliographic coupling network in the field of SC cybersecurity (VOSviewer)

2.1.5 Co-occurrence analysis

This section gives details of the analyses carried out on the co-occurrences of various keywords mentioned in the articles.

Author keywords analysis

A crucial aspect of bibliometric studies involves examining the co-occurrence of author keywords in the literature, which helps to identify the most focal themes within a research domain.

VOSviewer yielded a total of 5025 such keywords. To ensure a focused examination, a minimum occurrence threshold of 10 was applied, resulting in a list of 102 keywords with the most significant interrelationships. Fractional counting was utilized to weigh the co-occurrences, thereby equitably distributing emphasis across all contributing papers regardless of the number of keywords they contained.

The outcome of this analysis is presented in Table 2.5, displaying the top 15 author keywords arranged by their total link strength (TLS), a metric indicative of the keyword's

Rank	**Keyword**	**Occurences**	**TLS**
1	Cybersecurity	237	205
2	Machine Learning	224	202
3	Security	138	123
4	Cyber Security	122	105
5	Blockchain	113	102
6	Internet of Things	118	102
7	Anomaly Detection	118	102
8	Deep Learning	106	92
9	IoT	99	88
10	Threat Detection	105	87
12	Intrusion Detection	75	66
12	Industry 4.0	74	64
13	Network Security	61	50
14	Artificial Intelligence	57	49
15	Industrial Control Systems	49	46

Table 2.5: The top 15 keywords with the maximum occurrence values for author keywords.

centrality within the network. TLS values were calculated by summing the strengths of a keyword's link with all other interlinked terms.

This subsection presents a co-occurrence analysis of keywords cited by authors in cybersecurity research publications over the past six years. The search identified a corpus of 5025 author keywords from relevant research articles. To discern patterns and trends within this extensive dataset, a fractional counting method was employed, with a minimum occurrence threshold set at 10 for each keyword. This criterion was met by 102 keywords, indicating a concentrated focus on a subset of terms within the broader topic of supply chain cybersecurity.

Table 2.5 captures the essence of this research focus, presenting the top 15 keywords with the highest occurrence values. These terms represent the nexus of the discussion in supply chain cybersecurity literature. Notably, "Cybersecurity", "Security", and "Cyber Security" are omnipresent and expected terms within the research ambit. The term "Machine Learning" emerged as a central node, it boasts 224 occurrences, indicating its sheer volume in the literature, while its TLS of 202 reveals its interconnectedness with other research themes.

The computation of link strengths employs a fractional method where the interconnectivity between co-occurring keywords within an article is valued proportionally. For example, if ten keywords are interlinked within a publication, each pair's link strength is calculated as 1/10. The Total Link Strength (TLS) is reflective of the accumulated sum of such proportional link strengths across all articles, and it invariably materializes as a whole number. This happens because the aggregate of fractional link strengths for any keyword within an article always equals one, and thus the TLS becomes equivalent to the count of articles wherein the keyword is featured [55].

The curated list of 15 keywords delineates the thematic cores of the surveyed articles, with a notable emphasis on the detection and response to anomalies, threats, and intrusions through the application of machine learning, deep learning, and artificial intelligence technologies.

Figure 2.5, illustrates the co-occurrence network derived from this analysis, with author-defined cybersecurity keywords forming distinct clusters. For example, “Machine Learning”, “Classification”, “Malware Detection”, “Honeypot”, and “Data Mining” clusters within the network, are highlighted in orange. Meanwhile, “Cybersecurity”, “Threat Detection”, “Network Security”, “Vulnerabilities” and “Artificial Intelligence” are the key terms of the purple network. Finally, the red network includes “Security”, “Blockchain”, “Information Security”, “Internet of Things”, “Supply Chain”, “Big Data”, “Risk Management”, “Monitoring” and more.

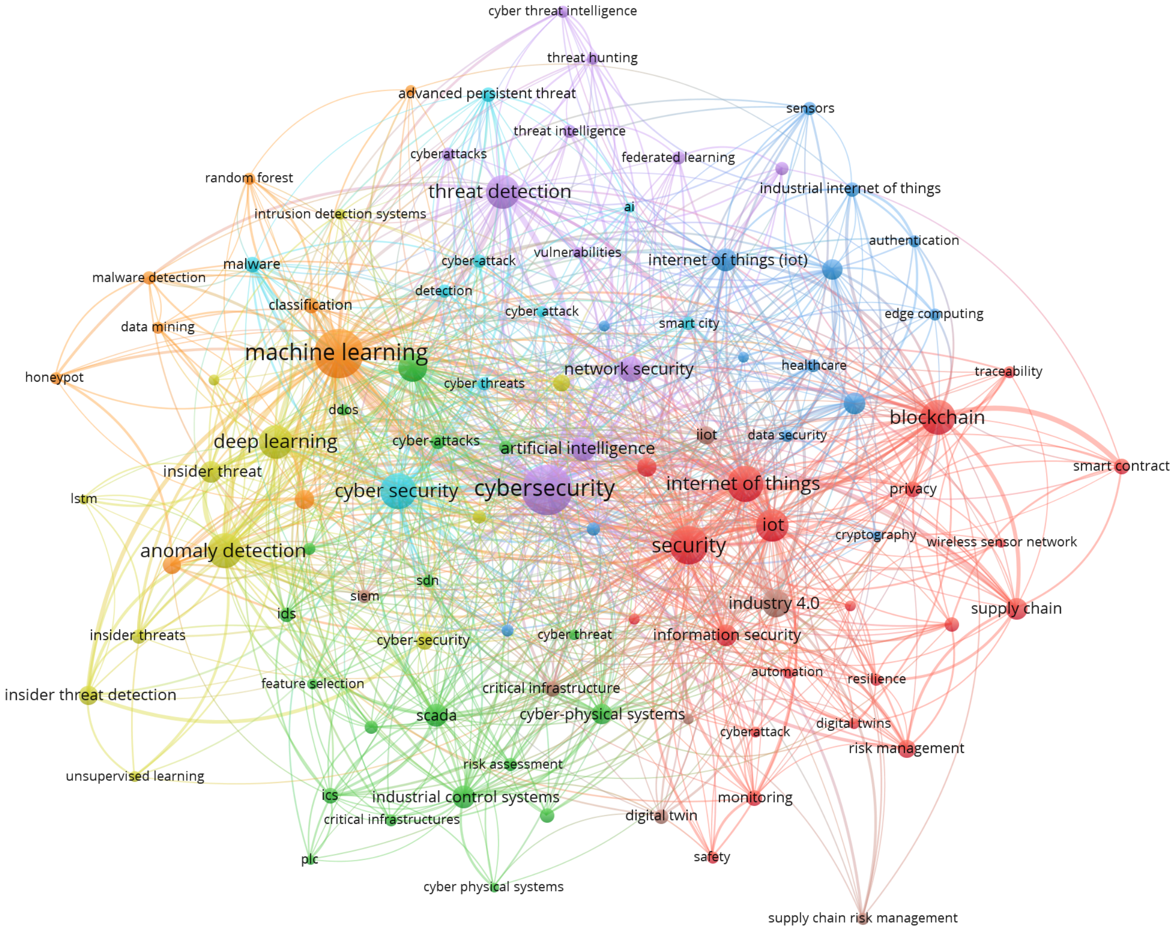

Figure 2.5: Co-occurrence network of author keywords in the field of SC cybersecurity (VOSviewer)

All keywords analysis

This subsection delves into the co-occurrence analysis of all keywords identified by VOSviewer from cybersecurity research publications over the past six years. The search amassed a total of 11982 keywords. Analyzed using the fractional counting method with a threshold set at a minimum of 10 occurrences per keyword, a filtered list of 438 keywords met this specified criterion. The top 15 keywords, distinguished by occurrence, are delineated in Table 2.6, illustrating the focal points within supply chain cybersecurity literature.

The findings reveal a shift in maximum co-occurrence values when the entire spectrum of identified keywords is considered. “Network Security” surfaced with 710 occurrences, in-

Rank	Keyword	Occurences	TLS
1	Network Security	710	707
2	Cybersecurity	544	543
3	Cyber Security	472	468
4	Internet of Things	434	432
5	Threat Detection	295	295
6	Machine Learning	293	292
7	Intrusion Detection	248	248
8	Learning Systems	201	201
9	Computer Crime	196	196
10	Cyber-attacks	194	194
12	Deep Learning	192	190
12	Embedded Systems	185	185
13	Supply Chains	183	183
14	Anomaly Detection	179	176
15	Information Management	175	175

Table 2.6: The top 15 keywords with the maximum occurrence values for all Keywords.

terlinking with 437 other keywords, accruing a TLS of 707. This TLS value is an aggregate of occurrences with each interrelated keyword that meets the occurrence threshold. Following this are the closely related “Cybersecurity” and “Cyber Security,” then “Internet of Things”, “Threat Detection”, and “Machine Learning”—the latter of which was previously ranked second in the author keyword analysis. Newly emerging terms include “Learning Systems”, “Computer Crime”, “Cyber-attacks”, “Embedded Systems”, “Supply Chain”, and “Information Management”, signaling an expansion of research interests.

There is a general trend observed where a higher frequency of occurrences is usually associated with an elevated TLS, indicating not only frequent mention but also extensive cross-connectivity within the body of literature.

Figure 2.6 offers a visual mapping of the keyword co-occurrence network, illustrating robust connections. “Network Security” is pivotal in a purple cluster, strongly associated with “Malware”, “Cyber Threat”, and “Cybersecurity”. In the green cluster, the interdependencies of “Threat Detection”, “Anomaly detection”, “Intrusion Detection” “Machine Learning”, “Deep Learning”, and “Learning Algorithms” are prominent. A red cluster highlights the interlinkage of “Internet of Things”, “Industry 4.0”, “Information Management”, “Security”, “Cryptography”, and “Blockchain”. The blue network, with smaller nodes, encompasses lesser-cited yet significant terms like “Industrial Control Systems”, “Embedded Systems”, “Cyber-attacks”, “Computer Crime”, and “Cyber-Physical Systems”. Finally, the yellow network, characterized by its smaller nodes, presents a cluster of terms such as “Supply Chains”, “Risk Management’ ”, “Risk Assessment”, and “Decision Making”, suggesting a less central but still relevant set of themes within the field.

The spread of keywords reveals the complexity and multifaceted nature of supply chain cybersecurity research, with a significant focus on technology-driven security measures, threat identification, and management strategies.

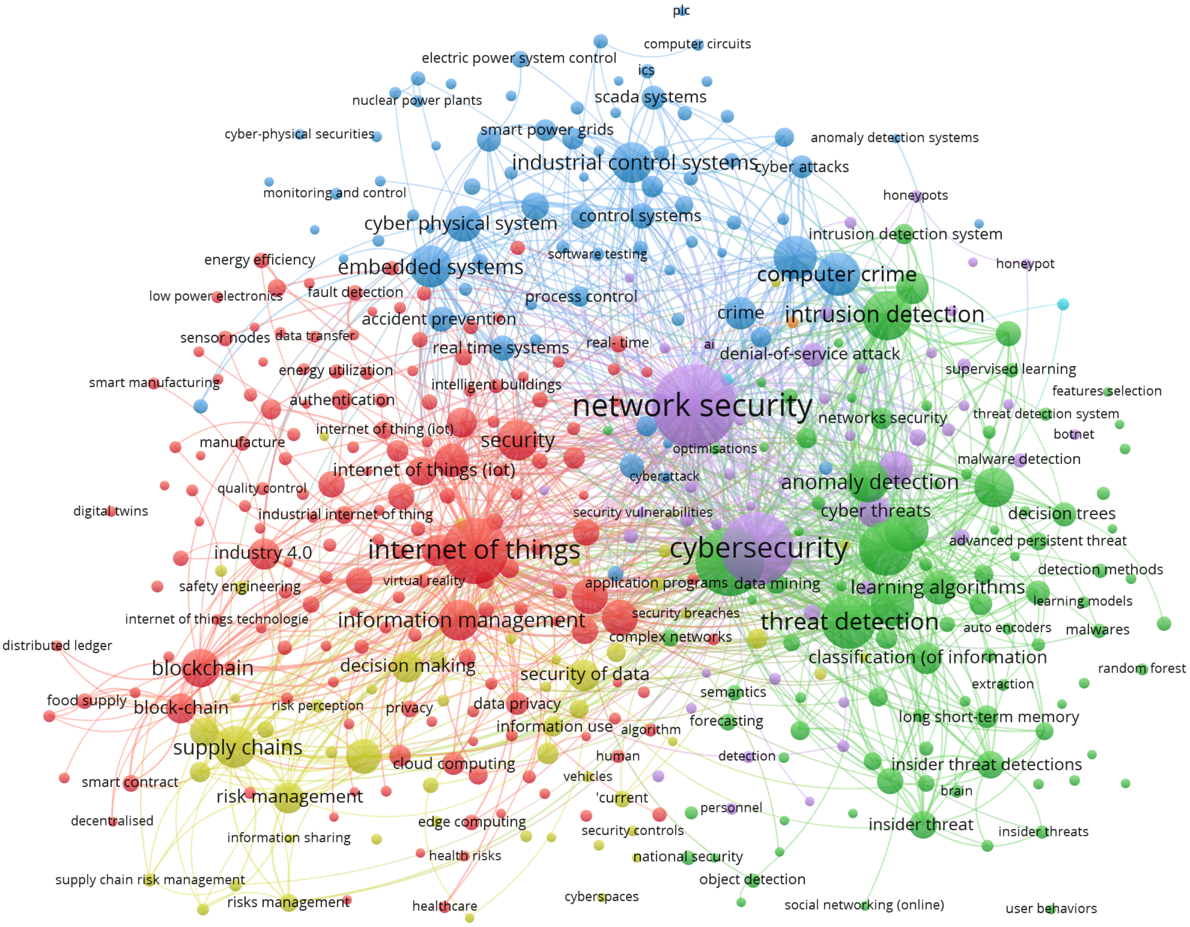

Figure 2.6: Co-occurrence network of all keywords in the field of SC cybersecurity (VOSviewer)

2.1.6 Conclusion

In constructing the literature review for this thesis, a rigorous bibliometric analysis was conducted to identify pivotal studies that encapsulate the multifaceted nature of supply chain cybersecurity.

Three documents emerged uniquely in response to a query centered on the inclusion of generative AI as a keyword (Q5 in Table 2.1), underscoring the niche application of generative AI and sophisticated machine learning paradigms in cybersecurity. Due to their specific focus and limited number, these documents were exempt from further selection criteria:

1. Wang et al.'s study on an insider threat detection framework harnessing digital twin technology and deep learning highlights innovative approaches to identifying internal security risks, which is crucial for safeguarding supply chain integrity [64].

2. FERRAG et al.'s exploration into the use of language models for IoT security emphasizes the transformative potential of AI in reinforcing data privacy and protection within the Internet of Things, a domain critical to supply chain security [23].

3. Alwahedi et al.'s comprehensive review of machine learning applications in IoT security provides foresight into the integration of generative AI with large language

models, charting the trajectory for future cybersecurity measures within supply chain systems [7].

Furthermore, the following documents have been selected for their scholarly significance as indicated by citation counts, bibliographic coupling, and co-citation metrics, reflecting their impact and pertinence.

In the domain of supply chain management, particularly concerning the integration of digital technology and Industry 4.0, the work of Ivanov et al., titled “The Impact of Digital Technology and Industry 4.0 on the Ripple Effect and Supply Chain Risk Analytics” [18] emerges as preeminent, securing the leading position in citation frequency, ranking sixth in co-citation, and second in bibliographic coupling. This document is instrumental in delineating the effects of digital advancements on supply chain vulnerabilities and the strategic management of associated risks.

Furthermore, Ivanov et al. contribute another foundational piece, “Digital Supply Chain Twins: Managing the Ripple Effect, Resilience, and Disruption Risks by Data-Driven Optimization, Simulation, and Visibility” which stands at the forefront of bibliographic coupling. This study delves into the utilization of digital twin technology, underscoring its significance in the real-time management and mitigation of supply chain disruptions.

The systematic review by Spieske et al., “Improving Supply Chain Resilience Through Industry 4.0: A Systematic Literature Review Under the Impressions of the COVID-19 Pandemic” [58] claims the fifth place in bibliographic coupling. The paper “Cyber Risk at the Edge: Current and Future Trends on Cyber Risk Analytics and Artificial Intelligence in the Industrial Internet of Things and Industry 4.0 Supply Chains” by Radanliev et al. positioned ninth in bibliographic coupling. It casts light on the evolving landscape of cyber risk analytics and the application of artificial intelligence at the edge of industrial IoT networks.

The research by Das et al., encapsulated in “Managing Disruptions and the Ripple Effect in Digital Supply Chains: Empirical Case Studies” [15] is recognized for its empirical contribution, placing tenth in bibliographic coupling. The document provides pragmatic insights into disruption management strategies within digital supply chains, offering empirical validation to the theoretical discourse.

Lastly, Mihai’s comprehensive survey “Digital Twins: A Survey on Enabling Technologies, Challenges, Trends and Future Prospects” [40] garners attention through the citation analysis normalization process. This paper’s enhanced visibility post-normalization underscores the burgeoning interest in digital twins and their transformative potential in supply chain cybersecurity.

The selected corpus deliver a thorough examination of critical themes such as AI’s role in cybersecurity and strategies for managing supply chain risks and disruptions. The inclusion of these documents ensures that the thesis is anchored in reputable and authoritative scholarly work.

2.2 Key Takeaways in Cyber Supply Chain Management

> Today and looking at the near future, the SC will be as good as the digital technology behind it.
>
> *Ivanov [33]*

Before the COVID-19 pandemic in early 2020, supply chains (SCs) were already dealing with challenges due to complex global setups and disruptions from natural disasters and political changes. The pandemic, however, exposed the vulnerabilities of SCs more dramatically than any previous event, affecting everything from supplier delays to unpredictable consumer demands and even full shutdowns of manufacturing due to strict health regulations [28, 52, 32]. The negative effects of these supply chain disruptions (SCDs) are severe, significantly harming overall corporate performance. They can lead to decreased sales and market share, lowered service quality, delays in delivery, and damaged reputation. As a response, there has been an increasing focus on strengthening supply chain resilience (SCRES), which seeks to quickly recover from such disruptions and, ideally, improve the supply chain's original performance [58].

Resilience is defined in the NIST Special Publication NIST SP 800-161r1, "Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations" as the capacity of a product, service, or supply chain to ensure an organization's ability to anticipate, prepare for, respond to, and recover from significant disruptions. This includes the ability to endure and rebound from intentional attacks, accidents, or natural threats and incidents [10].

2.2.1 Enhancing Cybersecurity in Supply Chain Management Through Generative AI

The utilization of Generative AI in cyber supply chain management, particularly in enhancing cybersecurity, is explored through a review of the three scholarly articles. These documents were specifically identified in response to a search query that included generative AI as a keyword (Q5 in Table 2.1). Each paper provides unique insights into the integration of Generative AI with current cybersecurity frameworks and highlights the potential for future advancements in this field.

In "DTITD: An Intelligent Insider Threat Detection Framework Based on Digital Twin and Self-Attention Based Deep Learning Models" by Wang et al., Generative AI, notably the Generative Pre-trained Transformer 2 (GPT-2), plays a pivotal role in augmenting cybersecurity measures. GPT-2 enhances insider threat detection systems by generating additional data to balance datasets, thus allowing for a more comprehensive understanding of insider threats. This model, in conjunction with BERT and DistilledTrans, improves the detection of anomalous patterns in user behavior, indicating potential security risks [64].

Another study, "Revolutionizing Cyber Threat Detection With Large Language Models: A Privacy-Preserving BERT-Based Lightweight Model for IoT/IIoT Devices" by Ferrag et al., employs Generative AI in the SecurityBERT architecture. This system leverages Transformer-based Large Language Models (LLMs) to analyze both structured and unstructured network data for security threats within IoT networks. Incorporating advanced encoding methods such as Privacy-Preserving Fixed-Length Encoding (PPFLE) and Byte-level Byte-Pair Encoding (BBPE), SecurityBERT processes sensitive data securely and efficiently, demonstrating superior accuracy and efficiency over traditional machine learning approaches in cyber threat detection [23].

Lastly, the article "Machine Learning Techniques for IoT Security: Current Research and Future Vision with Generative AI and Large Language Models" by Alwahedi et al. primarily discusses the use of Machine Learning (ML) in securing Internet of Things (IoT) environments through Intrusion Detection Systems (IDSs). It also speculates on the potential future applications of Generative AI and large language models in advancing cybersecurity measures in these environments [7].

These studies highlight the transformative potential of Generative AI in cybersecurity within the supply chain, showcasing its capability to enhance data quality, improve threat detection accuracy, and ensure secure data processing. The integration of these advanced AI models into cybersecurity frameworks presents a promising avenue for addressing the complex challenges of modern cybersecurity landscapes.

2.2.2 Managing Supply Chain Disruptions

In the study titled "Managing Disruptions and the Ripple Effect in Digital Supply Chains: Empirical Case Studies" by Das et al. [15], a survey highlighted significant concerns in supply chain management, focusing mainly on external and supplier risks. External risks mentioned include natural disasters such as fires and floods, severe weather affecting infrastructure, political instability, terrorism, and earthquakes at supplier locations.

The survey found that five out of nine companies were worried about supplier issues, such as changes in product quality or financial instability of suppliers. Other concerns raised included logistical problems during transport to sites or external warehouses, underestimated customer demand, and delays caused by machinery failures, production capacity issues, and heavy reliance on ocean freight. Respondents almost unanimously stated that disruptions in suppliers and demand directly caused interruptions in production capacity and consequently in deliveries to customers.

Participants were also asked about measures to prevent the ripple and bullwhip effects, particularly regarding the integration of suppliers into their risk management systems or the establishment of a risk management system with their suppliers. The risk manager from a company noted that they focus only on risks affecting their direct (first-tier) suppliers, as the next supplier level (second-tier) is often unknown, not disclosed by the first-tier suppliers, or not recorded in their system. He explained that monitoring all value-added levels is beyond their resources and that their purchasing conditions require first-tier suppliers to establish their own risk management systems. Moreover, he emphasized that taking over the risk management for suppliers would remove their accountability for any disruptions affecting production.

All respondents reported experiencing disruptions related to suppliers, production capacity, logistics, and demand, often with significant impacts. The study pointed out that responses did not specifically mention information disruptions or the ripple effects in supply chains. Additional problems included unexpected events like force majeure, tax changes, and regulatory shifts. The main causes of these disruptions were identified as dependence on a single supplier, insufficient production buffers, and poor visibility of data. Moreover, complex product specifications and changing customer requirements were also cited as reasons for product and technology disruptions. The risk manager also highlighted that not every component is equally critical in production, and risk management should focus particularly on supply-relevant components [15].

2.2.3 Technological Innovations in Supply Chain Resilience

The four-phase framework of Supply Chain Resilience (SCRES), developed by Hohenstein et al. [31], outlines the essential steps for managing disruptions in supply chains. The first phase, Readiness, involves preparing and setting up measures that help reduce the chance of disruptions and lessen their impacts. The next phase, Response, requires immediate action when a disruption is detected to minimize negative effects and stabilize the situation. The Recovery phase focuses on efforts to bring the supply chain back to its original performance level, addressing any lingering effects of the disruption. The final phase, Growth, aims to improve the supply chain's performance beyond its state before the disruption by incorporating lessons learned and adopting innovative practices, thus building a more resilient and competitive system. Most research on SCRES has traditionally emphasized the response phase, with less attention paid to the readiness and growth phases. However, the emergence of Industry 4.0 technologies has shifted focus toward the readiness phase, enabling more proactive strategies through advanced predictive analytics that offer insights into near-future scenarios [58].

The literature by Ivanov et al. on the impact of digital technology and Industry 4.0 on supply chain risk analytics and ripple effect reveals how proactive measures in SCM are significantly enhanced by these technologies. Their findings suggest that digital advancements increase demand responsiveness and capacity flexibility, potentially reducing the need for risk mitigation inventory, thus diminishing the ripple effect. Specifically, advancements in additive manufacturing contribute to shortened lead times, thereby optimizing inventory management practices. Furthermore, Industry 4.0 technologies, alongside Big Data Analytics (BDA) and Tracking and Tracing (T&T) systems, enhance the strategic planning of risk management infrastructures and enable a dynamic reconfiguration of resources during recovery stages.

During the reactive phase, these technologies facilitate unprecedented levels of data coordination and supply chain visibility, which are crucial for the effective activation and simulation of recovery policies. This enhancement in supply chain visibility allows for quicker deployment and more effective implementation of contingency plans developed during proactive stages. BDA, advanced T&T systems, and blockchain technology, in particular, are instrumental in tracing the origins of disruptions, observing disruption propagation (i.e., the ripple effect), and selecting stabilization actions based on a detailed understanding of available capacities and inventories [18].

The integration of digital technologies extends beyond manufacturing to include supplier networks, customer networks, and logistics service providers, aiming to enhance the overall flexibility of the supply chain in the face of disruptions. This comprehensive application underscores the importance of risk management across all supply chain actors, particularly in response to frequent incidents such as natural disasters or supplier disruptions. Understanding the sources and management processes of risks is crucial to harnessing the full potential of digital technologies [33].

However, the application of these technologies also introduces new challenges and risks. The use of Industry 4.0 and additive manufacturing can increase exposure to external risks due to the complexity they introduce, and while they reduce time and demand risks due to enhanced flexibility and shorter lead times, they also raise supply risks in scenarios where disruptions occur upstream and no intermediate inventory exists. The delivery process risks are also influenced by the capabilities of BDA to enhance supply chain visibility and forecast accuracy, thereby reducing demand risks and improving contingency plan quality, but increasing time risks due to heightened coordination complexity.

Digital technologies, particularly blockchain, can also play a pivotal role in reducing inefficiencies in risk management strategies by creating records of activities and data necessary for synchronized contingency planning. Moreover, decentralized control principles inherent in Industry 4.0 systems allow for a diversification of risks and a reduction in the need for structural supply chain redundancy through enhanced manufacturing flexibility [18].

The study by Das et al. categorically explores various digital technologies, evaluating their utility and implementation costs as rated by participants. BDA, Enterprise Resource Planning (ERP) systems, and tracking and tracing systems are highlighted as highly utilized across the companies surveyed, with BDA and ERP noted for their critical roles in strategic risk reduction and operational performance optimization. Industry 4.0 technologies and additive manufacturing, while not as widely adopted, are recognized for their potential to improve flexibility and reduce lead times, thereby enhancing supply chain responsiveness and reducing risk [15].

Additionally, the integration of these technologies facilitates significant improvements in real-time monitoring and supply chain visibility, which are paramount during both proactive and reactive phases of disruption management. For instance, advanced T&T systems combined with ERP systems bolster real-time data collection, crucial for managing disruptions effectively. The literature further underscores the utility of digital technologies in fostering a more resilient supply chain, particularly through their ability to enhance decision-making capabilities, support contingency planning, and facilitate rapid recovery from disruptions [15].

The literature indicates that designing a more resilient supply chain through these technologies is feasible, although it comes with increased information, external and supplier risks.

2.2.4 Advancements and Challenges in Digital Twin Technology for Supply Chain Management

In the evolving landscape of supply chain management (SCM), integrating analytics algorithms with optimization and simulation modeling is increasingly vital for competitive advantage. These technologies are transforming supply chains from fixed, physical systems into dynamic networks where firms dynamically allocate processes like supply, manufacturing, logistics, and sales.

The fusion of simulation and optimization allows for network optimization to minimize costs and enables dynamic policy analysis through simulation. This integrated approach is gaining popularity among supply chain managers for its enhanced decision-making capabilities.

A typical supply chain simulation-optimization model incorporates various elements for risk analysis, including GIS for site placement, and operational parameters like inventory control and production scheduling. These models can simulate disruptions using probability distributions and customize recovery strategies, providing a comprehensive view of the supply chain's health, which surpasses traditional models focused on single metrics.

Digital twins, as an extension of these models, mirror the physical supply chain in real-time, offering advanced risk analysis and operational management. They can quickly adapt to disruptions, testing recovery policies and adjusting contingency plans efficiently. The outputs from a digital twin can integrate with ERP systems or business intelligence tools to analyze disruption impacts and enhance operational performance [33].

In Mihai's study "Digital Twins: A Survey on Enabling Technologies, Challenges, Trends and Future Prospects", significant insights into digital twin technology are discussed, including the high costs and multidisciplinary nature of digital twins which complicate ROI calculations. The study also addresses the ethical and data quality challenges as digital twins expand into socio-technical systems. It emphasizes the importance of standardization and data security within digital twin ecosystems to ensure robust data governance and system integrity [40].

2.2.5 Review of NIST Guidelines for Cybersecurity Supply Chain Risk Management

The NIST Special Publication on Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations provides an in-depth framework for enhancing cybersecurity measures across supply chains. The document presents a well-structured approach to managing cybersecurity supply chain risks. It emphasizes the need to integrate C-SCRM across different levels of an organization and establish a formal program to effectively manage critical products, services, and suppliers. The publication emphasizes the importance of enterprises achieving a foundational level of maturity in key C-SCRM practices, as outlined in NIST standards, before progressing to more sophisticated capabilities. It advocates for adapting these practices to the specific context of the enterprise, such as risk profiles and available resources, to enhance the management of potential cybersecurity threats effectively.

Key practices highlighted include a thorough understanding of the enterprise's supply chain, close collaboration with critical suppliers, and the inclusion of these suppliers in resilience and improvement initiatives.

Additionally, the document discusses sustaining and enhancing practices that build upon the foundational ones. These practices involve establishing a threat-informed security program, utilizing third-party assessments to evaluate the security capabilities of critical suppliers, and continuous monitoring of the supply chain for changes in the risk profile. A formalized information-sharing function with other agencies and entities is recommended to enhance the enterprise's insights into threats and risks across the supply chain.

2.3 Conclusion

In the pursuit of enhancing supply chain resilience through cybersecurity, foundational practices identified by NIST provide a robust framework for addressing various supply chain vulnerabilities. Key practices underscore the necessity for a comprehensive understanding of the enterprise's supply chain, fostering close collaborations with critical suppliers, and actively involving these suppliers in resilience and improvement initiatives. This approach is crucial as it addresses multifaceted risks associated with information security, external threats, and supplier dependencies.

Furthermore, the literature reviewed indicates that designing a more resilient supply chain through these technologies is feasible, although it comes with increased information, external and supplier risk.

Based on the aforementioned considerations, this study intends to focus on the following foundational practices from NIST in my forthcoming research:

1. "Establish and begin using supplier risk-assessment processes on a prioritized basis (inclusive of criticality analysis, threat analysis, and vulnerability analysis) after the [FIPS 199] impact level has been defined." [10]

2. "Establish internal processes to validate that suppliers and service providers actively identify and disclose vulnerabilities in their products." [10]

N-tier supplier concentration significantly affects the resilience of supply chains. Up to 90% of critical nodes are found among N-tier suppliers, who remain mostly hidden due to limited visibility. This lack of transparency makes identifying and reducing risks more difficult both in the short and long term. N-tier visibility is crucial for mitigating structural risks by understanding supplier dependencies and developing alternative suppliers. It also enables broader visibility into high-risk suppliers, assesses risks based on geographic locations, ensures legal compliance by uncovering areas of non-compliance within the supplier network, and enhances crisis response capabilities by monitoring key N-tier suppliers for disruptions. Understanding the maturity of suppliers is essential to tailor engagements effectively and to identify latent risks or weaknesses within the supply chain network, facilitating transparency and swift, informed decision-making to mitigate risks [4].

The proposed solution involves developing a Generative AI Multi-Agent System designed to automate and enhance the N-tier mapping process through specific agent roles such as data collection, risk assessment, and compliance monitoring. The next steps in the research will involve defining and detailing the supplier N-tier process and integrating generative AI agents into this process.

Chapter 3

Enhancing Supply Chain Resilience through N-tier Mapping and Generative AI Integration

Supply chain resilience is the ability of a company's end-to-end supply chain to proactively sense, absorb, adapt to, and recover from disruptions. In the field of organizational resilience, having intelligent visibility is crucial. Firms with superior visibility are better prepared to withstand a variety of disruptions. Intelligent visibility encompasses both structural and dynamic visibility. Structural visibility can be likened to an X-ray, providing a static image of operations, while dynamic visibility is akin to a video, offering real-time monitoring and response capabilities.

One core requirement to establish resilience is the development of structural visibility. It reveals a company's operational state at a specific moment or over a period, aiding in identifying concealed issues. It encompasses conventional activities such as network mapping, risk assessment, network evaluations, and modeling [3].

Focusing on operational risk identification, the key to resilience is managing risk proactively. Companies can enhance their resilience by following a three-step process: initially assessing the current maturity level of their supply chain, subsequently creating a digital twin, and ultimately conducting resilience stress tests to determine a resilience score for their supplier network.

The literature review has already highlighted the importance of digital twins, which are instrumental for companies to establish a foundational level of structural visibility. By replicating the typical functions of a supply chain, a digital twin allows a company to employ advanced analytics to simulate and model scenarios of supply chain performance, as well as to conduct stress tests for risks and vulnerabilities. When creating a digital twin it is essential to include all known suppliers from Tier 1 to Tier N. Conducting a comprehensive N-tier mapping is a pivotal step, as it discloses the layers of suppliers and the materials they supply.

3.1 Enhancing Visibility through N-tier Mapping Processes

Creating visibility over the entire N-tier supply chain is crucial, as it allows organizations to identify potential bottlenecks and implement suitable measures to address them by understanding supplier concentration, single-sourcing, and dependencies and it also enhances visibility into high-risk suppliers and ESG impacts. It ensures legal compliance by uncovering non-compliance areas and adhering to new regulations. Moreover, it improves crisis response capabilities by ongoing monitoring key N-tier suppliers for disruptions and generating reports on GHG emissions and ESG risk baselines. Tailoring improvement engagement is facilitated by understanding the maturity of suppliers within the network. N-tier supply chain mapping can be executed through different approaches, each with its characteristics and challenges: questionnaire-driven, shipment data-driven and public data-driven methods.

3.1.1 Questionnaire-driven method

The questionnaire-driven method is a structured process that involves the identification of Tier 1-2 suppliers, followed by the distribution of detailed questionnaires. Once the responses are collected and analyzed, Tier 3 + N-tier suppliers are identified and mapped. The process then continues the N-tier mapping with further questionnaires and ends with the establishment of ongoing risk monitoring. Key inputs include Tier 1-2 supplier locations, internal manufacturing and logistics sites, supply paths, BoM data, and purchase orders and invoices. The outcome is a comprehensive mapping of the supply chain up to the N-tier level, providing live risk monitoring at both the supplier and material levels, including indirect Tier 1 suppliers. The method is acclaimed for its high accuracy and the ability to uncover physical locations and establish mapping at the material/SKU level. However, these benefits come at a cost, including the extensive time required to complete the process, low automation, and a generally low response rate without thorough supplier education. The timeline largely depends on supplier response time.

3.1.2 Shipment data-driven method

The shipment data-driven method begins by prioritizing categories and Tier 1 suppliers, creating a list of suppliers per category for discovery. Next, a list of Tier 1 suppliers, including material delivered and supplementary data, is compiled for each category and sent to an external analyst partner. The third-party partner then conducts Tier 2 discovery. Following this, validation and any necessary additional research are performed. The review and validation of Tier 2 results are conducted in-house using market and supply chain information. Subsequently, the third-party partner refines Tier 2 and proceeds with Tier 3 discovery. This step is followed by a final review and in-house validation, ensuring completeness, fine-tuning, and accuracy of the results. Finally, the validated data is uploaded to a risk and alerts platform. This balanced approach, which combines manual and automated methods, ensures accuracy and speed and includes the identification of physical supplier locations. However, it is limited to direct suppliers and requires comprehensive knowledge of the value chain. The process provides N-tier mapping for selected categories, logistics flow identification, and live risk monitoring at the material level.

3.1.3 Public data-driven method

The public data-driven method primarily utilizes publicly available data. It starts by prioritizing categories and Tier 1 suppliers, creating a list of suppliers per category for discovery. A third-party partner then undertakes a comprehensive N-tier discovery using public records data partners. This is followed by in-house data cleansing. Subsequently, a final in-house review and validation are conducted to ensure completeness, fine-tuning, and accuracy of the results. The validated data is then uploaded to a risk and alerts platform. This highly automated and efficient process can identify both direct and indirect suppliers and detect sanctioned entities. However, it is limited to legal entities and headquarters levels only, and does not provide insights into physical sites.

As previously highlighted, the questionnaire-driven approach remains distinctive among mapping methods due to its accuracy and comprehensive detail. However, despite its precision, this approach faces notable challenges in efficiency, primarily caused by the significant time required for supplier education and data collection. Moreover, its dependency on manual processes leads to minimal automation, hindering the efficient implementation and updating of N-tier mapping, thereby limiting the overall effectiveness of this method.

3.1.4 Integration of N-tier Supply Chain Mapping Approaches

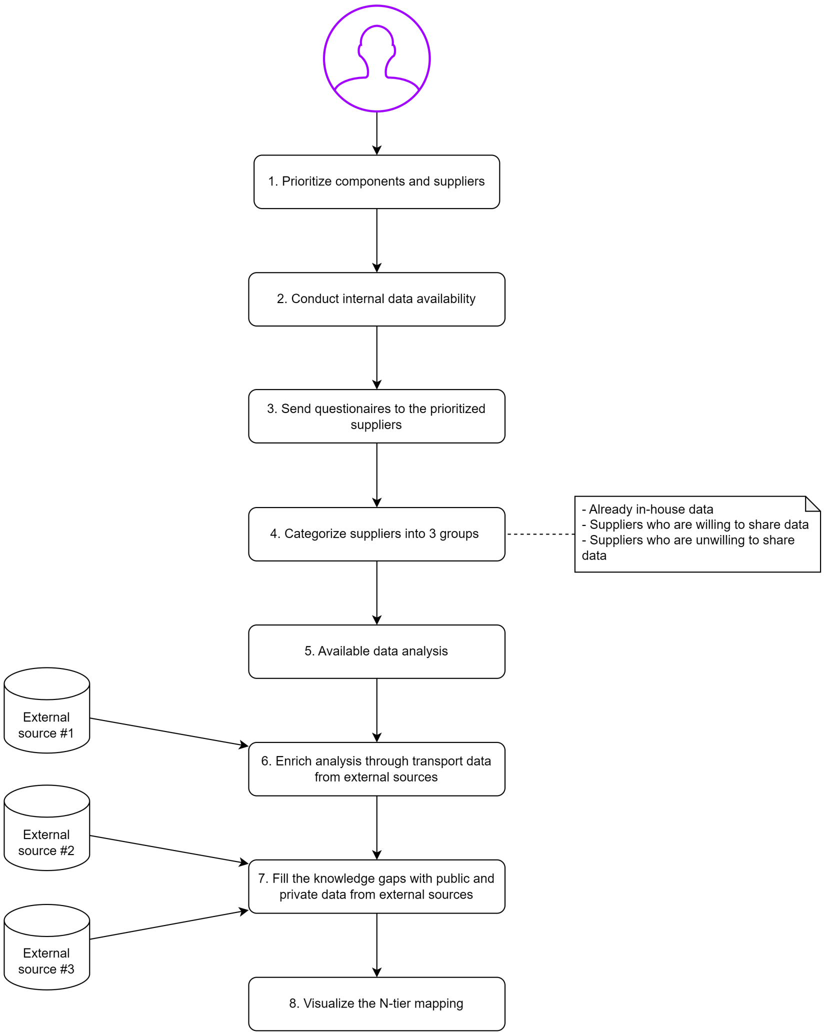
1. Prioritize components and suppliers
2. Conduct internal data availability
3. Send questionaires to the prioritized suppliers
4. Categorize suppliers into 3 groups
- Already in-house data
- Suppliers who are willing to share data
- Suppliers who are unwilling to share data
5. Available data analysis
External source #1
6. Enrich analysis through transport data from external sources
External source #2
7. Fill the knowledge gaps with public and private data from external sources
External source #3
8. Visualize the N-tier mapping

Figure 3.1: Overview of the N-tier mapping process. Own illustration.

To capitalize on the inherent strengths of these approaches while addressing their limitations, this study aims to develop a multi-agent system that combines the questionnaire-driven, shipment data-driven, and public data-driven approaches, leveraging generative AI capabilities to automate the process. This system will focus on automating data collection, enhancing response rates, and expediting questionnaire analysis, particularly within the questionnaire-driven methodology. By integrating these approaches, the aim is to create a highly efficient and precise N-tier mapping process.

The process of N-tier mapping involves systematically collecting, analyzing, and visualizing supply chain data through several critical steps, as illustrated in Figure 3.1. Later in this chapter, we will delve into some steps in more detail.

Initially, the process requires the identification and prioritization of key components and suppliers critical to the supply chain. This prioritization ensures that efforts are focused on the most significant areas.

Following the prioritization, an internal data availability assessment is conducted. This assessment involves reviewing the existing organizational data to ascertain what information is already available and identifying any data gaps.

Suppliers are then categorized into three groups: those with sufficient in-house data, those willing to share their data, and those unwilling to share data, including non-respondents. This categorization helps streamline subsequent data analysis efforts.

To supplement the internal data, questionnaires are distributed to the suppliers. These questionnaires are designed to collect specific information not available internally. Later in this chapter, we will delve into this step in more detail.

The next phase involves a thorough analysis of the collected data to identify N-tier suppliers and potential risks within the supply chain. This analysis is further enriched by integrating transport data from external sources, which provides additional context and addresses gaps that internal data and supplier responses may not cover. Later in this chapter, we will delve into this step in more detail.

To address any remaining knowledge gaps, the process incorporates both public and private data from external databases and resources. This data is then combined into a bespoke database, which serves as a foundation for in-depth analysis. This holistic strategy guarantees a more thorough and precise comprehension of the supply chain dynamics, leveraging multiple data sources to construct a singular, comprehensive repository for analysis.

Finally, the analyzed data is visualized in an N-tier mapping, illustrating the complex relationships and dependencies across different tiers of suppliers. This visualization is crucial for stakeholders to comprehend the intricacies of the supply chain effectively.

Within the step of sending out questionnaires Figure 3.1, the subsequent process involves the utilization of a pre-established list of prioritized suppliers and materials categorized for discovery. The approach commences with gathering Tier 1 data by conducting thorough research on supplier names, locations, and products, which results in a structured approach for supplier engagement. Following this, a questionnaire is created, involving the generation, refinement, and completion of a pilot questionnaire, yielding a validated survey template ready for supplier engagement.

The next phase involves engaging Tier 1 and Tier 2 suppliers by initially dispatching questionnaires to Tier 1 suppliers while preparing for engagement with Tier 2 suppliers. This step results in the collection of Tier 2 and subsequent Tier data, along with confirmation of Tier 2 engagement readiness. After collecting the data, it is analyzed to identify gaps and acquire Tier 2 discovery information, leading to enhanced insights and a refined template for Tier 2 suppliers. Surveys are then issued to Tier 2 suppliers, replicating the analysis process previously conducted for Tier 1 suppliers, culminating in the acquisition of Tier 3 and subsequent Tier data. The final step in this series involves the verification, unification, and preparation of questionnaire data for shipment data enrichment.

The enrichment of supplier findings is supplemented by transport data and market research. This process begins with filling product knowledge gaps through external research and validating the list of critical subcomponents with the client. Afterward, a thorough overview of the subcomponents market is developed, detailing suppliers and main export countries. The identification of UNSPSC codes for the subcomponents within the scope follows, along with the selection of N-tier components/materials for further research. Unstructured transport data is then collected from an external partner, cleansed, and transformed. This data is filtered based on predefined criteria and validated for completeness using market overview information. The results are then validated with the client, and the analysis is fine-tuned based on client feedback. The outcome of this process is a comprehensive overview of the N-tier components market.

3.2 Strategic Integration of LLMs in N-tier Process Optimization

The intricate process of N-tier mapping underscores the necessity for advancements in Large Language Models (LLMs). Currently, LLMs are primarily utilized in zero-shot mode, generating output sequentially without the opportunity for revision. This method is analogous to asking someone to compose an essay without the ability to backtrack, aiming for a high-quality result. Despite this challenge, LLMs perform remarkably well. However, an agent workflow introduces the concept of iteration, enabling the LLM to revisit and refine its output multiple times. Through this iterative workflow, AI can achieve significantly improved results compared to a single-pass approach [46].

Before delving into specific patterns, it is essential to understand the common elements utilized across these designs. One fundamental component in building agents is the prompting techniques applied to the language model. The typical method for creating prompts involves incorporating intermediate reasoning steps, known as Chain of Thought (CoT) prompting. This technique allows the model to solve complex problems incrementally, generating more accurate answers compared to zero-shot prompting. Another technique, Reasoning and Acting (ReAct) prompting, extends CoT by including actionable steps within an environment [57].

Further techniques, such as self-reflection, introduce verbal reinforcement to the model, encouraging self-assessment of previous steps and enhancing the model's ability to generate accurate responses [47, 57]. Additionally, tool use empowers LLMs to leverage external resources, enhancing their functional capabilities and effectiveness in diverse environments [48, 57]. The use of personas in initial prompts also helps maintain focus on specific problem domains, improving the quality of the generated output. For maintaining contextual coherence and enhancing learning from interactions memory capabilities, both short-term and long-term, are crucial. Furthermore, Finite State Machines (FSM) provide a computational model for designing and analyzing agent behaviors, allowing transitions between states based on previous outputs or specific conditions. [57].

To further leverage LLMs, they have been employed as reasoning engines for autonomous agents. Autonomous agents are AI programs capable of performing complex tasks with a degree of independence. LLM-based agents have recently shown remarkable potential in reasoning and planning, aligning closely with human expectations for autonomous agents that can perceive, decide, and act in response to their environment. The foundational principles of such agent development, as outlined in Sniffin's article, are expressed through three design patterns [57].

Firsty, the ReAct agent specializes in reasoning and action handling within its environment. In this design pattern, different ReAct prompts are implemented as specific states in a FSM. This approach ensures the agent's responses are consistent and relevant to the current context or task at hand. Each state can represent various properties, including a prompt for the model and a handler for mapping application logic to and from the model. The main states in the ReAct pattern are: Thought—addresses the problem given the previous actions and determines the next step to take, Act—determines the correct tool to use and the correct input for that tool, and Observe—summarizes the behavior from the action to the memory. The advantages of this agent implementation strategy include enhanced predictability, isolation of tasks from other states, straightforward troubleshooting, and the simplicity of integrating new states. However, potential issues may arise, such as a propensity for the agent to become trapped in loops or to deviate from the initial request, losing focus on the original objective [57].

Secondly, the Task-Planner agent extends the ReAct agent by introducing a planning step. This agent defines a concrete plan on what needs to be done and attempts to work through that plan in multiple steps. The plan consists of tasks where each task is an isolated piece of work. Similar to the ReAct agent, the design for this agent can use an FSM as the basis of its implementation. The planning step occurs as a new state, and the action state pops tasks from the stack and observes the output from the tools. The advantage of this pattern is that work is planned upfront, which can help reduce the chance of getting stuck in a loop, although this is not guaranteed. Initial mistakes in the plan can cause errors throughout the tasks, necessitating backtracking and generating new tasks. Such issues can be costly, so planning should be limited to tasks that are easily predictable [49, 57].

Lastly, multi-agent orchestration involves the collaboration of multiple AI agents, each specializing in specific problems. This design allows agents to split up tasks, discuss, and debate ideas to develop better solutions than a single agent would. One approach to managing complex tasks is by separating responsibilities and introducing a reasoning step for communication among agents. By enabling agents to communicate, they can delegate work through the orchestration of tasks, similar to a delegation-like pattern. Instead of having a single agent handle everything, agents are defined to specialize in solving specific problems with different implementations. An orchestrator supervises and routes tasks between agents to achieve the best-desired output. By separating and abstracting communication, agents do not need to understand how a task is solved [50, 57].

Furthermore, the Microsoft article introduced the concept of MicroAgents [54]. When developing an AI personal assistant capable of managing a diverse array of services—such as location, calendar, email, banking, shopping, travel, and weather—the MicroAgent pattern offers a compelling architectural strategy. This approach diverges significantly from the monolithic design, which integrates all functionalities and plugins within a singular agent model. In a monolithic system, the agent must determine which of potentially thousands of services to invoke, complicating the provision of nuanced system instructions. In contrast, the MicroAgent pattern partitions the system by functional domains, associating each microagent with a specific service.

The adoption of microservices in this context confers numerous benefits. Firstly, ease of maintenance is achieved through simplified feature enhancement and validation processes. This modularity allows for more straightforward updates and improvements. Secondly, reliability is enhanced by improved fault isolation and diagnostics, which help to identify and rectify issues more efficiently. Lastly, deployment agility is significantly improved, allowing for more frequent updates and scalable deployment practices.

3.2.1 Designing Large Language Model-Based Multi-Agent Systems

The development of LLM-based Multi-Agent(LLM-MA) systems necessitates a comprehensive grasp of their overarching structure which is dissected into four key aspects by Guo et al. [29]. These include the interface between agents and their environment, the profiling of individual agents to define their roles and specialties, the communication protocols that enable agents to collaborate and share information, and the methods through which agents acquire and refine their capabilities to perform complex tasks effectively.

Agents-Environment Interface

The operational environment specifies the contexts in which LLM-MA systems are deployed, encompassing domains such as software development, gaming, financial markets, and social behavior modeling. We delineate three primary categories of interfaces within LLM-MA systems:

- The **Sandbox** interface denotes simulated or virtual environments enabling agents to engage and experiment freely. This is prominently utilized in domains such as software development and gaming.
- Conversely, the **Physical** interface pertains to real-world environments where agents interact with physical entities and adhere to real-world constraints, as exemplified by robotic tasks.
- The **None** interface encompasses scenarios devoid of any specific external environment, focusing solely on inter-agent communication and consensus-building activities.

The interaction between agents and their operational environments significantly influences their behaviors and decision-making processes. LLM-based agents perceive and act within these environments, receiving feedback that informs strategy adjustments over time. The Agents-Environment Interface is pivotal for agents to comprehend their surroundings, make informed decisions, and learn from the repercussions of their actions. This iterative process of perception, action, and feedback is essential for the continuous refinement of agent strategies and behaviors.

Agents Profiling

In LLM-MA systems, agents are characterized by their traits, actions, and skills, which are tailored to meet specific goals. Across various systems, agents assume distinct roles with comprehensive descriptions that encompass characteristics, capabilities, behaviors, and constraints. For example, in software development, agents could take on roles such as product managers and engineers, each with specific responsibilities and expertise that guide the development process. Similarly, in a debating platform, agents might be designated as proponents, opponents, or judges, each with unique functions and strategies to fulfill their roles effectively. These profiles are crucial for defining agents' interactions and effectiveness within their respective environments. Regarding the methods of agent profiling, they are categorized into three types: **Pre-defined**, **Model-Generated**, and **Data-Derived**. Pre-defined profiles are explicitly defined by system designers, Model-Generated profiles are created by models such as large language models, and Data-Derived profiles are constructed based on pre-existing datasets.

Agents Communication

Communication between agents in LLM-MA systems is the critical infrastructure supporting collective intelligence. This communication is dissected from three perspectives: Communication Paradigms, Communication Structure, and Communication Content. The three paradigms for agent communication are **Cooperative**, **Debate**, and **Competitive**.

Cooperative agents work together towards shared goals, typically exchanging information to enhance collective solutions. The Debate paradigm involves agents engaging in argumentative interactions to reach a consensus or a more refined solution. Competitive agents work towards their own goals, which might conflict with those of other agents. The communication structure can be **layered**, **decentralized**, **centralized**, or involve a **shared message pool**. Layered communication is hierarchically structured, decentralized communication operates on a peer-to-peer network, centralized communication involves a central coordinating agent, and a shared message pool involves agents publishing and subscribing to messages to enhance efficiency. The communication content is typically in the form of **text**, varying widely depending on the application.

Agents Capabilities Acquisition

For enabling agents to learn and evolve dynamically, the acquisition of capabilities in LLM-MA systems is crucial. Feedback is a critical component of this process, helping agents understand the impact of their actions and adapt to complex problems. Feedback can come from the environment, other agents, humans, or, in some cases, may not be provided at all. Agents adjust to complex problems through memory, self-evolution, and dynamic generation. Memory involves storing information from previous interactions to enhance current actions. Self-evolution allows agents to dynamically modify their goals and strategies based on feedback or communication logs. Dynamic generation enables the system to create new agents on the fly to address current needs and challenges. The increasing complexity of managing various agents has made agents' orchestration a pivotal challenge in scaling up LLM-MA systems.

3.2.2 Conceptual Framework for Implementing LLM-MA Systems in N-tier Mapping Processes

The LLM-MA system (Figure 3.2) integrates multiple autonomous agents to collaboratively perform N-tier mapping within a defined operational environment. This system primarily functions within a sandbox environment, a virtual space created by humans where agents can freely interact with each other and with databases essential for their tasks. This environment is critical for simulating the interactions and processes required for effective N-tier mapping.

The design and selection of agents within the LLM-MA are tailored to meet the unique requirements of the N-tier mapping process. This ensures that each agent can effectively contribute to achieving N-tier mapping goals previously discussed. Initially, the N-tier mapping process was carefully reviewed and broken down into smaller, specific tasks. Each of these tasks was then assigned to agents designed to handle them effectively. For example, Value Chain Analysts are solely focused on identifying key components and suppliers.

In the design phase of these agents, key considerations included their ability to operate concurrently and their adaptability for reuse. Having multiple agents work simultaneously speeds up the mapping process allowing several suppliers to be analyzed at the same time. Moreover, agents are built to perform specific tasks, such as data analysis or questionnaire distribution, and they can be easily adapted or slightly altered for similar roles in different projects.

Each agent in the system is created using pre-defined profiles. These profiles outline the expected behaviors, actions, and skills needed for each agent to effectively perform its role. This ensures consistency and reliability, helping the system function smoothly and efficiently across various mapping projects.

In this multi-agent system, teamwork is essential. By combining agents with different roles and perspectives, the system uses their varied abilities to work together and achieve the common goal of efficient and accurate N-tier mapping.

Stakeholders relationship in the MA system

The system's architecture is a mixed structure incorporating both centralized and sequential elements to efficiently manage complex processes and ensure effective communication among multiple agents. This design choice enhances the system's functionality and scalability.

At the core is an orchestration agent that supervises and routes tasks among agents. The orchestration agent acts as a control hub, simplifying the scaling of agent behaviors and ensuring that all parts of the system align with the overall objectives. This orchestration agent employs cooperative communication with all agents and can receive feedback from human users.

Centralized coordination allows the system to efficiently handle an increase in the number of tasks or complexity without significant redesign. The orchestration agent streamlines the process of task delegation. Instead of each agent determining who to communicate with, the orchestrator assigns tasks, ensuring that each agent is utilized according to its specialty. By overseeing and evaluating the output of each subprocess, the orchestration center can determine its adequacy.

The subprocesses within the system, such as the questionnaire distribution, employ a sequential structure where tasks are handled one after the other in a linear progression by various agents. This sequential approach ensures that each task is fully completed before the next one commences, which significantly reduces the risk of errors and inconsistencies that might occur with concurrent processing.

This structure allows for a streamlined troubleshooting process, as it becomes simpler to identify issues when tasks are handled in order. Moreover, this setup facilitates a feedback mechanism where subsequent agents can review and provide feedback on the outputs from previous agents. This feedback management ensures that each stage of the process meets the necessary quality benchmarks before moving on to the next.

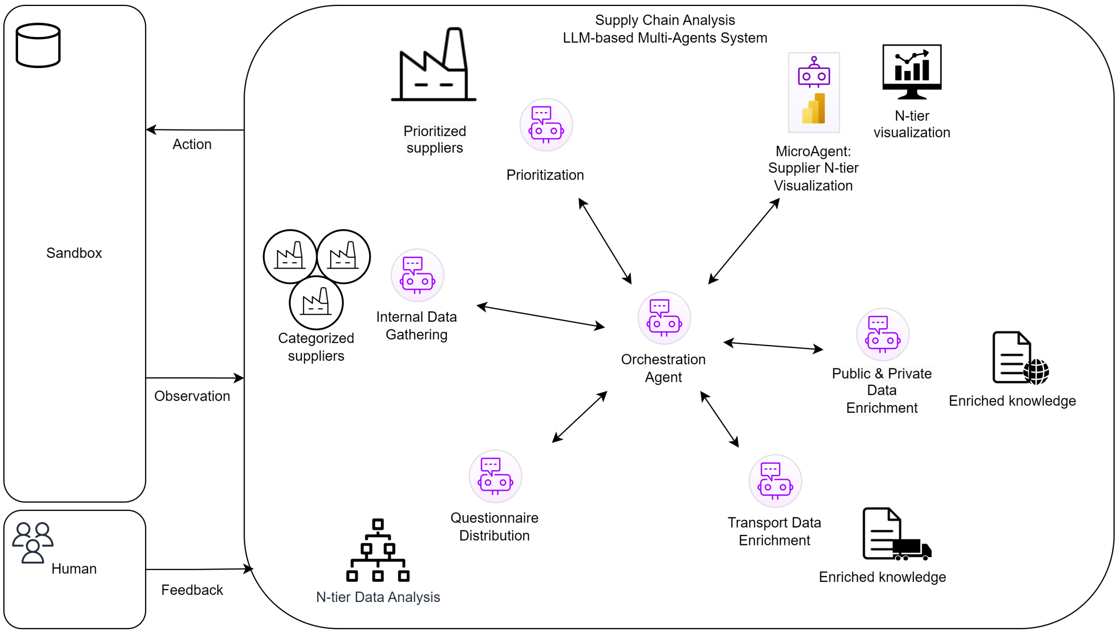

Figure 3.2: Schematic Overview of the LLM-MA System for N-tier Mapping Process.

Prioritization Subprocess

The orchestration agent initiates the Prioritization subprocess (Figure 3.3) for components and suppliers. Agents, acting as Value Chain Analysts, receive the value chain as input data and prioritize critical components and suppliers. This process involves debate among agents to refine the prioritized list, supplemented by human feedback to enhance capability development. Agents use a memory module to store and retrieve valuable information from previous interactions.

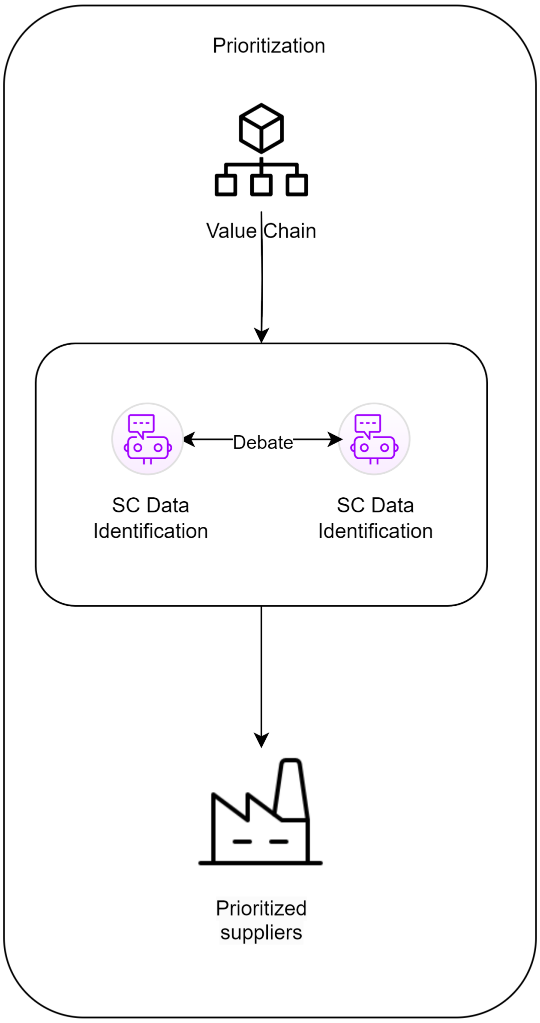

Figure 3.3: Subprocess for Prioritization of Components and Suppliers in the LLM-MA System.

Data Gathering and Analysis

Upon human verification of the prioritization outcome, the orchestration agent triggers the agent whose role is Data Analyst to check the availability of internal data on prioritized suppliers (Figure 3.2). The agent gathers this data from various sources, using self-evolution to adapt its strategy dynamically. The process outputs a categorized list of suppliers, distinguishing those requiring further questionnaire-based data collection from those with existing in-house data.

Questionnaire Distribution Subprocess

For each supplier identified for questionnaire-based data collection, the orchestration agent triggers the relevant subprocess (Figure 3.4). The Data Analyst Agent gathers necessary supplier information from different sources and communicates with Questionnaire Generation Agents, who act as Cybersecurity Consultants to debate and refine the questionnaires. The Analyst Agent utilizes self-evolution to adapt its strategy dynamically and gets feedback from the Questionnaire Generation Agents. Human feedback is implemented for the questionnaire generation process to ensure high-quality outputs. The Questionnaire Generation Agents use memory modules for behavior adjustment based on previous feedback and interactions.

The finalized questionnaires, approved by human consultants, are sent to suppliers via an Email Sender Microagent. Suppliers can complete the questionnaires with assistance from a Chatbot Agent providing real-time assistance to suppliers who may have difficulties or questions when completing the questionnaires. This agent can guide suppliers through the questionnaire, clarify questions, provide examples, and ensure that the data collected is accurate and complete. By doing so, it reduces the need for extensive supplier education and streamlines the data collection process. The chatbot incorporates feedback from both suppliers and professionals to improve its performance. To adapt to its tasks, the agent utilizes a memory module, allowing it to adjust its behavior based on past interactions. If a supplier does not respond, their data will be analyzed using transport and public data analysis.

The collected data is then forwarded to the Task Planner Agent, who functions as a Data Analyst. This agent formulates a plan to unify and analyze the data for subsequent steps. During the analysis, the agent identifies previously hidden suppliers and employs self-evolution to refine its methods. The orchestration agent then restarts the internal data-gathering process with these newly identified suppliers, ensuring a comprehensive and dynamic approach to data management and analysis.

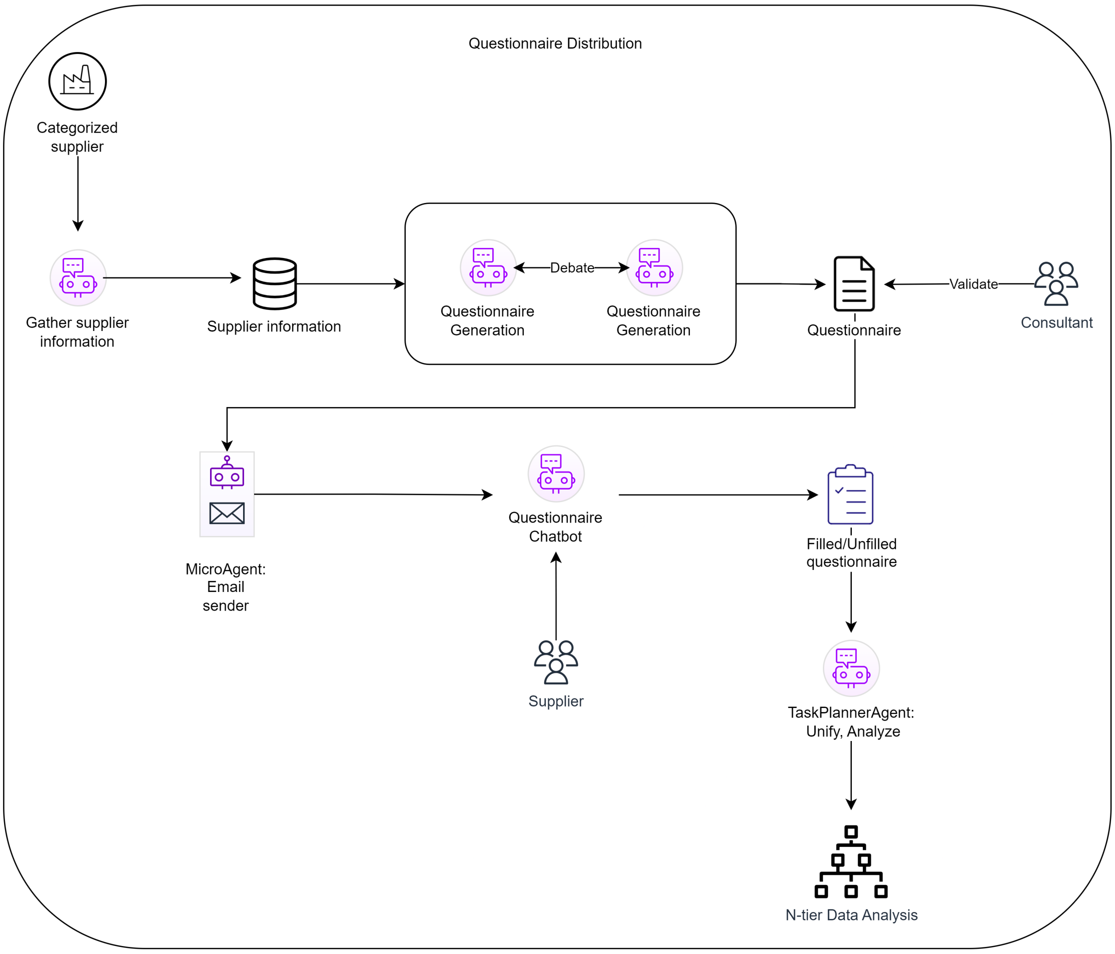

Figure 3.4: Subprocess of the Questionnaire Distribution in the LLM-MA System.

Enrichment with Transport Data

To enhance the supplier analysis, the orchestration agent initiates the transport data enrichment process (Figure 3.5). First, a Researcher Agent collects information about the supplier's products from external sources, using self-evolution to adapt to its task. If the gathered data is insufficient, feedback from the next agent helps refine the process. The next agent identifies critical subcomponents based on the collected data, leveraging a memory module to store and retrieve valuable past interactions. This agent also receives feedback from the client, ensuring that its analysis is aligned with client needs. The output of this process is a comprehensive overview of the critical subcomponents.

Another agent then identifies the UNSPSC codes for these subcomponents and gathers unstructured transport data for the selected N-tier components and materials from external sources. Acting as an Analyst, this agent uses self-evolution and receives feedback from the subsequent agent. Next, a task planner agent designs a plan to clean, transform, filter, and validate the collected data, using client feedback and self-evolution to refine its approach.

The final task planner agent in the subprocess enriches the supplier analysis with the cleaned and filtered transport data. This agent also employs self-evolution to continuously improve its performance.

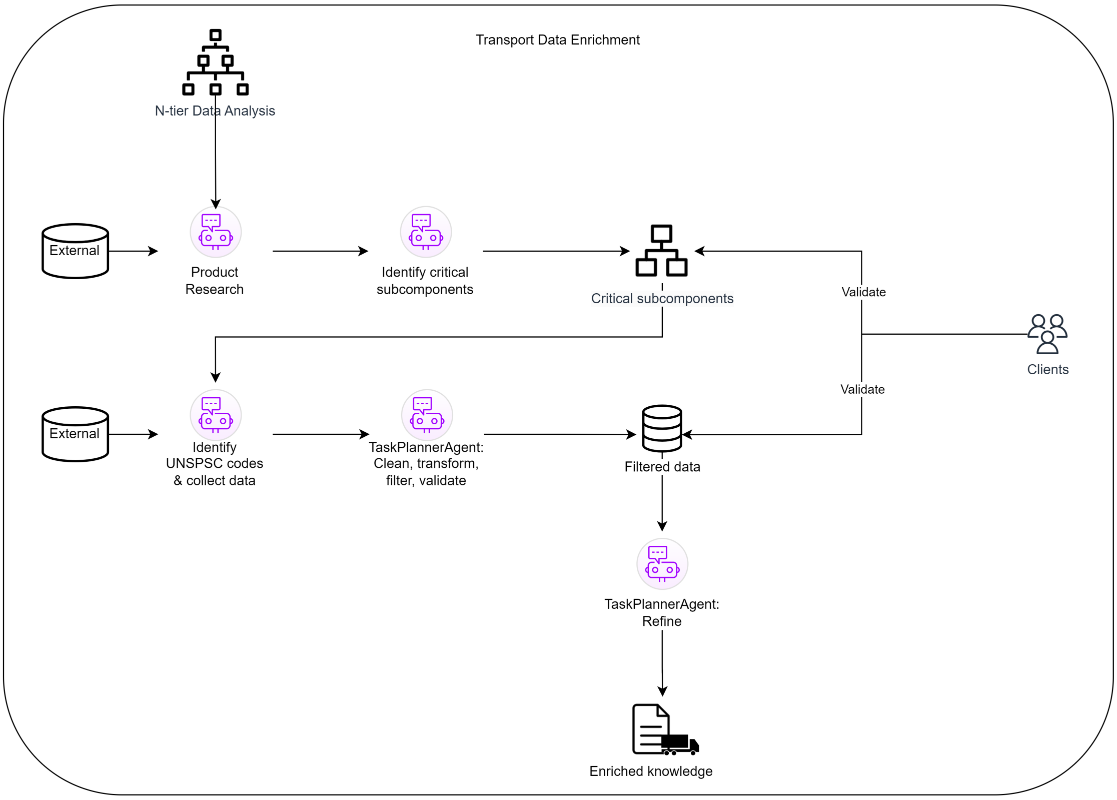

Figure 3.5: Subprocess for Enrichment with Transport Data in the LLM-MA System.

Enrichment with Public Data

To fill the identified knowledge gaps, the orchestration agent initiates the public data enrichment process (Figure 3.6). Public data is sourced from predefined sources and combined into a comprehensive database. An Analyst Agent filters and analyzes this data to fill any gaps, leveraging self-evolution and feedback from the orchestration agent. The output is an enriched understanding of the N-tier mapping.

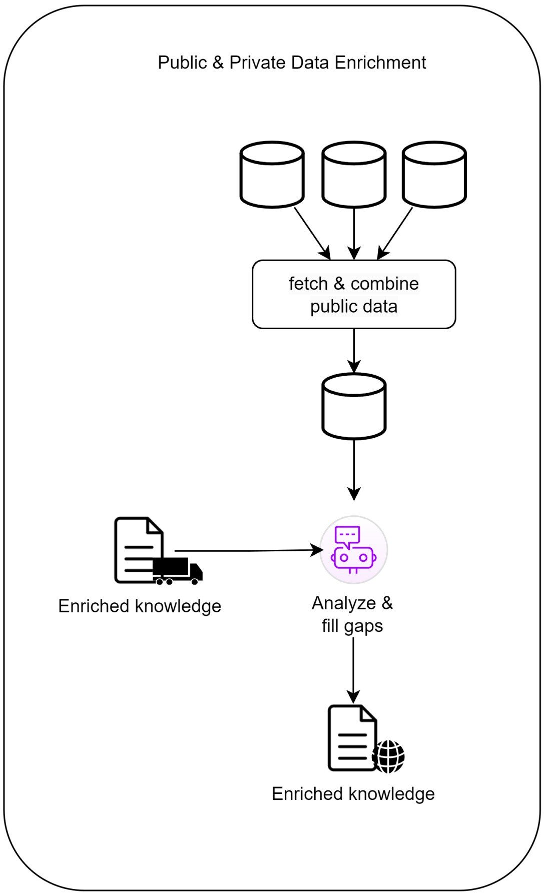

Figure 3.6: Subprocess for Enrichment with Public Data in the LLM-MA System.

Visualization

The final step involves visualizing the N-tier mapping results (Figure 3.2). The orchestration agent transmits the results to a Visualization MicroAgent, which uses PowerBI to create visual representations of the N-tier mapping.

Chapter 4

Experiments

In the study, several advanced artificial intelligence models were used, namely GPT-4 by OpenAI (2024), and Claude 3 and Llama 3 by Corcel (2024). The aim was to compare their performance in complex tasks. To mimic a multi-agent system, the outputs of other agents were communicated to each model through user prompts. This allowed the models to act as if it were part of a multi-agent system.

The experiments focused on specific agents within the multi-agent system. The tasks performed by several specialized agents within the multi-agent system were evaluated. These included the Internal Data Gathering Agent, Questionnaire Generation Agent, Questionnaire Analyst Agent, Orchestration Agent, Public Data Analyst Agent, and Prioritization Agent. The selection of these particular agents for testing was strategic, as their tasks encapsulate the full range of functions performed by other types of agents within the system. This comprehensive testing approach ensures that the evaluation covers all critical aspects of the multi-agent system's operational capabilities. Agents with similar tasks were assumed to behave similarly, such as analysis and information gathering agents. Agents related to chatbot functions, email sending, or visualization tasks were not included in this phase of the investigation due to limited resources.

Given the sensitive nature of the data involved in typical N-tier systems, and to prevent any breaches of corporate secrets, artificially generated data were used for the simulations. This ensured that the experiments did not compromise any proprietary or sensitive business information. Due to compliance issues, the transport data enrichment component was not included in the experiments.

A key part of the examinations was prompt engineering. This involves carefully designing input prompts that guide the AI models in performing specific tasks. Each agent's task was divided into a clear, step-by-step workflow in the setup. This helps produce a clear and directed response from the AI agents and improves predictability.

The prompting strategy was based on zero-shot learning principles. In language models like GPT-4, zero-shot learning means the model's ability to carry out tasks without having seen specific examples before. We used a method similar to zero-shot ReAct prompting, where the system prompts are structured to include both "thought" and "action" steps.

In the process of refining our prompting strategy, adjustments to the prompts and inputs were iteratively made until the desired outcomes were achieved. This involved a series of trials and errors to identify the most effective phrasing and structuring of the prompts. The iterations continued until the responses from the AI models met the predetermined criteria for accomplishing the task. It is important to note that only the final versions of these optimized prompts are presented in this study.

This structured and detailed methodology provides insights into the potential applications and limitations of using AI in multi-agent systems, contributing significantly to AI research and its practical implementations in cyber security industry-specific technologies.

4.1 Prompt Design

The primary goal of the prompt design was to guide the GEN AI agents in performing complex tasks, such as analyzing value chains, prioritizing components and suppliers, etc., which they were not specifically trained for. Since the agents, including models like GPT-4, Claude 3, and Llama 3, are general AI models without specific training on the nuanced cybersecurity tasks, the prompts had to be meticulously structured to achieve the desired outputs.

Facilitating a systematic, step-by-step process ensures comprehensive analysis and clear output generation. This "think before act" methodology is vital in helping general AI models apply their capabilities to specialized tasks effectively. By guiding the AI through a structured thought process before action, the prompts help overcome the models' inherent limitations in task-specific training.

The prompts were broken down into distinct components, each designed with specific considerations:

1. Persona definition

   Defining the agent's role at the beginning sets clear expectations, which is essential when working with general AI models lacking specific training for the tasks. This initial clarity helps focus the AI's efforts on the desired analytical goals.

2. Step-by-Step Instructions

   GEN AI can manage the analysis more effectively by breaking the tasks into smaller, sequential steps despite not being specifically trained. This segmented approach ensures thorough coverage of each task component.

3. Interactive Components

   Incorporating interactive feedback codes such as "CONTINUE" into the prompts is crucial due to the output limitations of the models. These codes enable iterative processing, allowing the agent to handle complex tasks progressively and deliver outputs that are too detailed to be generated in one session. This iterative process is key to ensuring that each part of the task is fully addressed before moving on.

For a practical demonstration of the principles applied in this study's prompt design and AI interactions, you can explore a detailed example here.

4.2 Procedure

In the initialization phase, multiple Large Language Model (LLM) agents are generated through distinct conversations. These agents are instructed via detailed prompts to comprehend their designated personas and the specific tasks associated with the N-tier mapping process.

During this initialization phase, each agent undergoes a comprehensive briefing, which includes an in-depth understanding of their roles and the expectations set for their task of the N-tier mapping process. This preparation ensures that each agent is equipped with the necessary contextual knowledge and directives to perform their assigned tasks effectively.

Subsequently, in the interaction phase, each agent is tasked with performing their designated responsibilities based on the provided input data. This process is iterated multiple times across various LLM models, including ChatGPT, LLama 3, and Claude 3. The zero-shot ReAct prompts are designed to emulate a "think before act" methodology, thereby enhancing the quality and accuracy of the task execution. This reflective practice encourages the agents to process the input data thoroughly before initiating their actions, leading to improved outcomes.

Following the interaction phase, a qualitative analysis is conducted using the enhanced RACCCA framework (Relevance, Accuracy, Completeness, Clarity, Coherence, and Appropriateness, with an added emphasis on Consistency) [62]. This framework is designed to comprehensively assess the performance of each Large Language Model (LLM) agent based on criteria crucial for the successful execution of tasks. This is essential, as the effectiveness of LLMs in complex tasks such as N-tier mapping relies heavily on the subtleties of language understanding and decision-making which are not easily quantifiable through traditional metrics. In order to provide a structured and quantifiable measure of the performance of each Large Language Model (LLM) agent a three-level classification method is applied to each criterion. This grading scale ranges from 1 to 3, where 1 indicates poor performance, 2 indicates adequate performance, and 3 represents excellent performance. Two independent reviewers evaluate the agents' performance based on these criteria across three iterations of each task for each model. The evaluation process involves both reviewers assessing the same outputs to determine two average grades for each criterion for each model. These average grades are calculated from the scores given by the reviewers across the three iterations. This method ensures that the evaluation reflects the general performance of the models rather than the outcomes from any single instance.

1. Relevance: Determines if the responses directly address the core of the tasks assigned. This assesses whether the agents' outputs are on point with the expected tasks or if they deviate from the set objectives.
    (a) Grade 1: The agent's responses significantly deviate from the task objectives, failing to address the core issues or questions posed.
    (b) Grade 2: The responses generally address the main points of the task but may include some irrelevant details or omit minor relevant aspects.
    (c) Grade 3: The agent's responses are directly on point, comprehensively addressing all aspects of the task without any deviations.
2. Accuracy: Focuses on the factual correctness of the agents' responses. It is critical in scenarios where the outputs influence decision-making processes or other agents within the system.
    (a) Grade 1: The responses contain significant factual errors, misleading information, or incorrect interpretations that could adversely affect decision-making or the integrity of the system.
    (b) Grade 2: The responses are mostly accurate but may contain slight inaccuracies that do not fundamentally undermine their utility.
    (c) Grade 3: The responses are factually correct and provide accurate information relevant to the task at hand.
3. Completeness: Checks whether all parts of the instructions are fully addressed in the agents' responses. This ensures that no aspect of the task is overlooked, which is vital for the integrity of the N-tier mapping process.

(a) Grade 1: The responses are incomplete, addressing only a subset of the required elements of the task.

(b) Grade 2: The responses cover most of the task requirements but may miss some details that are not critical for the overall comprehension and outcome of the task.

(c) Grade 3: The responses are comprehensive, addressing all components of the task thoroughly and leaving no aspect unattended.

4. Clarity: Evaluates how easily the responses can be understood. Clear communication is essential for ensuring that subsequent actions based on these responses are correctly implemented.

   (a) Grade 1: The responses are confusing, poorly structured, or use complex language that makes comprehension difficult.

   (b) Grade 2: The responses are generally clear but might require some effort to understand certain parts or could be expressed more succinctly.

   (c) Grade 3: The responses are articulated in a clear, concise, and well-structured manner, facilitating easy understanding and implementation.

5. Coherence: Measures how logically the ideas are presented and connected within the responses. Coherence helps in maintaining a smooth flow of information that aligns with the logical structure of the task.

   (a) Grade 1: The responses are disjointed or illogical, with ideas and statements that do not connect well, disrupting the flow of information.

   (b) Grade 2: The responses have a logical flow but may display minor inconsistencies that slightly disrupt the narrative or argument.

   (c) Grade 3: The responses are logically organized and coherent, with a seamless flow of ideas that builds a strong, understandable narrative.

6. Appropriateness: Assesses the tone and style of the responses, ensuring they are suitable for the context in which they are used. This includes checking if the language and formalities align with the professional standards expected in an operational setting.

   (a) Grade 1: The tone, style, or language used in the responses is inappropriate for the context, potentially leading to misunderstandings or a lack of professionalism.

   (b) Grade 2: The responses are mostly appropriate, but there might be occasional lapses in tone or style that do not seriously affect the overall appropriateness.

   (c) Grade 3: The responses perfectly match the expected tone, style, and professional language suitable for the context.

7. Consistency: Newly added to enhance the framework, this criterion measures the consistency of the agent's responses when given the same input with the same prompt. Consistency ensures that the agent's performance is reliable and stable, producing similar outputs across multiple iterations under identical conditions.

   (a) Grade 1: The agent produces varied and inconsistent responses when faced with the same task under identical conditions, leading to unpredictable outputs.

(b) Grade 2: The responses are somewhat consistent, with some variation that does not significantly impact the reliability of the agent.

(c) Grade 3: The agent delivers highly consistent outputs, ensuring reliable and predictable responses across multiple iterations.

To quantify the agreement between the two reviewers and ensure the reliability of the evaluations, the standard deviation and Cohen's kappa are calculated for each criterion of each agent's task. Standard Deviation is calculated to determine the variability of the review scores around their average. A standard deviation of zero indicates no variability, suggesting perfect agreement between reviewers. Cohen's Kappa measures the agreement between the two reviewers adjusted for the agreement that could occur by chance. A Cohen's kappa of 1 indicates perfect agreement.

4.3 Results

4.3.1 Orchestration Agent

The orchestration agent in the system serves as a central coordinator, managing and directing tasks among various agents within a mixed centralized and sequential architecture. It ensures optimal outcomes by facilitating cooperative communication among all agents.

Predefined Criteria

The Orchestration Agent must possess the capability to discern the origins of the output, pinpointing the specific source agent, and subsequently channel it to the appropriate destination agent.

Input Data

The orchestration agent's input comes from the other agents from the system, which have been subjected to rigorous testing across diverse models.

In one instance, the orchestration agent processed a list of suppliers, segmented into two categories: those requiring a questionnaire and those for which a questionnaire was deemed unnecessary. The objective of this routing process was to direct the initial segment to the agent responsible for generating questionnaires, while the latter segment was routed to the agent tasked with enriching transport data.

Final Agent Prompt

You are an orchestration agent in a multi-agent system for N-tier mapping. Here is the process that you should follow.

1. Initiate the Prioritization subprocess. Call the Value Chain Analyst Agents to prioritize components and suppliers. Await output.

2. If the output starts with "Output from Value Chain Analyst:", verify the prioritization with human feedback and proceed to Internal Data Gathering.

3. Call the Internal Data Gathering Agent to check the availability of internal data on prioritized suppliers. Await a categorized list of suppliers.

4. If the output starts with "Output from Internal Data Gathering Agent:", identify suppliers requiring questionnaire-based data collection and trigger the Supplier Information Gathering Agent in the Questionnaire Distribution Subprocess. For those who are not categorized for questionnaire, initiate the transport data enrichment process. Call the Product Researcher Agent for external data collection.

5. If the output starts with "Output from Questionnaire Data Analyst:", restart the internal data-gathering process with the newly identified suppliers. For the others, initiate the transport data enrichment process. Call the Product Researcher Agent for external data collection.

6. If the output starts with "Output from Transport Data Analyst Agent:", initiate the public data enrichment process. Call the Analyst Agent to filter and analyze public data. Await enriched N-tier mapping.

7. If the output starts with "Output from Public Data Analyst Agent:", proceed to Visualization. Transmit the results to the Visualization MicroAgent for the creation of visual representations using PowerBI.

Note: Always ensure that the output is verified and that the correct subsequent agent is called based on the output received. Maintain communication with human users for feedback and verification throughout the process.

Evaluation Grading for the Orchestration Agent

1. **Relevance:**
   - (a) **Grade 1:** The agent's responses are significantly off-target, focusing on unrelated tasks or misinterpreting the flow of operations between agents.
   - (b) **Grade 2:** The orchestration agent generally manages tasks correctly but might occasionally route information to the wrong agent or misinterpret data categories.
   - (c) **Grade 3:** The orchestration agent accurately identifies and directs tasks and data between agents, perfectly aligning with the procedural flow outlined in the prompt.

2. **Accuracy:**
   - (a) **Grade 1:** The agent incorrectly identifies the sources of outputs or routes data inaccurately, significantly disrupting the system's operation.
   - (b) **Grade 2:** The agent mostly routes data correctly but may occasionally make errors in data source identification or destination, which are promptly corrected.
   - (c) **Grade 3:** The agent flawlessly recognizes the origins of data and routes it without error, ensuring accurate operations throughout the system.

3. **Completeness:**
   - (a) **Grade 1:** The agent addresses only some parts of the routing process, frequently missing critical steps or ignoring specific agent outputs.
   - (b) **Grade 2:** The agent covers most routing tasks but may overlook minor details that do not severely impact the overall operation.
   - (c) **Grade 3:** The agent comprehensively manages all routing tasks, ensuring that no part of the process is overlooked or mishandled.

4. **Clarity:**
   - (a) **Grade 1:** The agent's communications are unclear or confusing, leading to frequent misinterpretations and errors in task execution by other agents.
   - (b) **Grade 2:** The agent's instructions are generally clear but might contain ambiguities that occasionally need clarification.
   - (c) **Grade 3:** The agent communicates precisely and clearly, making the routing and task management processes easily understandable for all involved agents.

5. **Coherence:**
   - (a) **Grade 1:** The agent's responses are inconsistent and disconnected, resulting in a fragmented and inefficient task management process.
   - (b) **Grade 2:** The agent generally maintains a logical flow in its operations, though there are minor inconsistencies that slightly disrupt the process.

(c) **Grade 3:** The agent exhibits a perfectly coherent response pattern, with logical and well-connected steps that ensure smooth operations across the system.

6. **Appropriateness:**

   (a) **Grade 1:** The agent uses a style or tone unsuited for an operational environment, potentially causing confusion or misinterpretation among other agents.

   (b) **Grade 2:** The agent's tone and style are largely appropriate, though occasional lapses do not significantly impact its effectiveness.

   (c) **Grade 3:** The agent consistently uses a professional tone and style that is perfectly suited for the operational context, enhancing understanding and efficiency.

7. **Consistency:**

   (a) **Grade 1:** The agent provides varied and unpredictable responses when processing identical inputs under the same conditions, leading to unreliable outcomes.

   (b) **Grade 2:** The agent shows some consistency in handling tasks, with only slight variations that do not majorly affect the reliability of its performance.

   (c) **Grade 3:** The agent delivers highly consistent and reliable outputs, ensuring stable and predictable responses in every iteration.

Outcome

Table 4.1: Qualitative Analysis of Orchestration Agent Performance

Criteria	**GPT-4**		**Claude 3**		**Llama 3**	
	Review 1	Review 2	Review 1	Review 2	Review 1	Review 2
Relevance	3	3	3	3	3	3
Accuracy	3	3	3	3	3	3
Completeness	3	3	3	3	3	3
Clarity	3	3	3	3	3	3
Coherence	3	3	3	3	3	3
Appropriateness	3	3	3	3	3	3
Consistency	3	3	3	3	3	3
Total Average	3	3	3	3	3	3

In the conducted trials, the Orchestration Agent adeptly directed the received outputs to the appropriate destination agent, demonstrating consistent proficiency across all tested scenarios. Notably, the agent exhibited accurate routing capabilities when evaluated with each of the three specified models, totaling nine tests conducted, each achieving a flawless success rate of 100

As shown in Table 4.1, the qualitative analysis of the Orchestration Agent's performance reveals that all three models—GPT-4, Claude 3, and Llama 3—scored highly across all criteria. Each model demonstrated high relevance, accuracy, completeness, clarity, coherence, appropriateness, and consistency in their responses.

Inter-Rater Reliability in Orchestration Agent Performance Evaluation

The evaluation of the orchestration agent's performance using standard deviations and Cohen's Kappa coefficients reveals exceptionally high inter-rater reliability across all criteria for the models GPT-4, Claude 3, and Llama 3. Both metrics indicate perfect agreement among the reviewers, with standard deviations consistently at zero, suggesting no variance in the evaluations. Similarly, Cohen's Kappa values are uniformly at 1, indicating perfect agreement with what could be expected by chance.

Table 4.2: Standard Deviations of Orchestration Agent Performance

Criterion	**GPT-4**	**Claude 3**	**Llama 3**	**All**
Relevance	0	0	0	0
Accuracy	0	0	0	0
Completeness	0	0	0	0
Clarity	0	0	0	0
Coherence	0	0	0	0
Appropriateness	0	0	0	0
Consistency	0	0	0	0

Table 4.3: Cohen's Kappa of Orchestration Agent Performance

Criterion	**GPT-4**	**Claude 3**	**Llama 3**	**All**
Relevance	1	1	1	1
Accuracy	1	1	1	1
Completeness	1	1	1	1
Clarity	1	1	1	1
Coherence	1	1	1	1
Appropriateness	1	1	1	1
Consistency	1	1	1	1

4.3.2 Prioritization Agent

The Value Chain Analyst agents within the orchestration framework are tasked with analyzing the value chain to identify and prioritize critical components and suppliers. They engage in a deliberative process to refine their assessments and produce a prioritized list, which is facilitated by the orchestration agent's initiation of the Prioritization subprocess.

Predefined Criteria

The criteria of these agents is to come up with a list of suppliers that the process should focus on. The created list should include what are the high-, medium- and low-priority components and who are the high- and medium-priority suppliers of the supply chain.

Input Data

The input data is structured as a hierarchical representation of a bank's value chain, detailing the essential products and processes organized into three primary categories. These categories encompass both the foundational elements and client-facing services that underpin the bank's operations and customer interactions.

1. Facilitory Products: This first category captures the core operational supports and services necessary for the bank's day-to-day functions and long-term financial management. This includes various products that are crucial for transaction handling, financial planning, and risk management.
2. Client Products: The second category focuses on the direct offerings to customers, crucial for ensuring client satisfaction and engagement. It covers a range of services and operations that facilitate customer transactions and interactions with the bank through various payment and banking platforms.
3. Retail NL Tech Core Process: The third category details the essential processes involved in managing customer relationships and service delivery within the retail banking sector in the Netherlands. It covers everything from customer onboarding to handling disputes and regulatory compliance.

The input also emphasizes the interconnections between these categories, illustrating how each component supports and enhances the others to ensure efficient, secure, and customer-focused banking operations. This structured approach not only helps in understanding the individual parts of the bank's operations but also their collective role in achieving operational excellence and customer satisfaction.

Final Initial Agent Prompt

You are a Value Chain Analyst Agent, tasked with analyzing the value chain and prioritizing components and suppliers. Your role is crucial in ensuring that the most critical elements are identified and prioritized effectively. Follow the tasks outlined below:

1. Receive and Analyze Input Data: - Obtain the value chain input data from the Orchestration Agent. - Perform an initial analysis to identify key components and suppliers within the value chain.

2. Prioritization Process: - Prioritize the identified components and suppliers based on their importance and impact on the value chain. - Use established criteria and metrics to rank the components and suppliers.

Output Handling: - Once the prioritization is refined and agreed upon, generate a final list of prioritized components and suppliers. - Ensure the output is comprehensive and well-documented for the Orchestration Agent to proceed with the next steps.

Your diligent analysis and collaborative efforts are vital to the success of the N-tier mapping process, ensuring that the most critical components and suppliers are accurately prioritized for further scrutiny and action.

Value chain: <AGENT_INPUT>

Complete one task at a time. At each step complete some parts of the tasks and summarize your findings at the end of your answer. If you receive the code: CONTINUE then continue with completing some parts of your task. Once you are done with all the stages then output the keyword: DONE

Final Debate Agent Prompt

You are a Value Chain Analyst Agent, tasked with analyzing the value chain and prioritizing components and suppliers. Your role is crucial in ensuring that the most critical elements are identified and prioritized effectively.

Your task is to: - Engage in constructive debates with other Value Chain Analyst Agents to discuss your prioritization findings. - Consider alternative perspectives and data points presented by other agents. - Refine the prioritized list through collaborative discussions and consensus-building.

At each step of the debate if you feel an agreement has been reached, then write the keyword: DONE

Evaluation Grading for the Prioritization Agent

1. **Relevance:**
   (a) **Grade 1:** The agent does not focus on the crucial components or suppliers, instead analyzing irrelevant parts of the value chain.
   (b) **Grade 2:** The agent generally prioritizes important components and suppliers but may include non-essential items in its analysis or overlook some critical elements.
   (c) **Grade 3:** The agent perfectly identifies and prioritizes all critical components and suppliers, directly aligning with the goals set out in its tasks.

2. **Accuracy:**
   (a) **Grade 1:** The agent makes significant errors in the identification or ranking of components and suppliers, potentially leading to flawed prioritization.
   (b) **Grade 2:** The agent is mostly accurate in its rankings but may have minor discrepancies that do not generally affect the overall outcome of the prioritization.
   (c) **Grade 3:** The agent accurately assesses and ranks all components and suppliers according to their importance and impact, without any errors.

3. **Completeness:**
   (a) **Grade 1:** The agent provides a partial list that omits several key components or suppliers.
   (b) **Grade 2:** The agent produces a comprehensive list but may miss some details that do not critically impact the prioritization outcome.
   (c) **Grade 3:** The agent's output is thorough, including a well-documented and detailed prioritization of all relevant components and suppliers.

4. **Clarity:**
   (a) **Grade 1:** The output from the agent is unclear or poorly organized, making it difficult to discern the prioritization or reasoning.
   (b) **Grade 2:** The agent's output is generally clear with some areas that might require further clarification or could be more succinct.
   (c) **Grade 3:** The agent presents its findings in a clear, concise, and well-structured manner, facilitating easy understanding and subsequent decision-making.

5. **Coherence:**
   (a) **Grade 1:** The prioritization logic is flawed or illogical, with significant gaps in the reasoning process.
   (b) **Grade 2:** The prioritization is logical for the most part, with minor inconsistencies that may slightly detract from the overall logic.
   (c) **Grade 3:** The agent's prioritization is logically sound, with coherent reasoning that effectively supports the ranking of components and suppliers.

6. **Appropriateness:**
   (a) **Grade 1:** The tone or style of the agent's output is inappropriate for the professional setting, potentially leading to misunderstandings.

(b) **Grade 2:** The overall tone and style are suitable, with occasional lapses that do not significantly detract from its professionalism.

(c) **Grade 3:** The agent consistently maintains a professional tone and style that is entirely appropriate for the context and enhances the usability of its output.

7. **Consistency:**

   (a) **Grade 1:** Outputs vary significantly when dealing with similar data under the same conditions, leading to unreliable prioritizations.

   (b) **Grade 2:** There is some variation in the agent's outputs, but these do not generally undermine the reliability of its prioritizations.

   (c) **Grade 3:** The agent provides highly consistent outputs, ensuring reliable and predictable prioritization across different instances.

Outcome

Table 4.4: Qualitative Analysis of Prioritization Agent Performance

Criteria	**GPT-4**		**Claude 3**		**Llama 3**	
	Review 1	Review 2	Review 1	Review 2	Review 1	Review 2
Relevance	3	2.67	1.67	1.3	1	1
Accuracy	2.67	2.3	1.67	2	1	1
Completeness	3	3	1.3	1.3	1	1.3
Clarity	3	3	3	3	3	3
Coherence	3	3	1.67	2	1	1.3
Appropriateness	3	3	3	3	3	3
Consistency	3	2	1	1	1	1
Total Average	2.95	2.71	1.9	1.94	1.57	1.65

In the preliminary assessment (Table 4.4) of the performance of AI models, Claude is characterized by its highly inconsistent outputs. It is notable that on several occasions, Claude failed to provide any response during the initial interaction phase. An illustrative example of this is when Claude erroneously produced a list titled "TOP 10 suppliers," which deviated significantly from the specified task. Llama's contributions were consistently off-topic, demonstrating a lack of relevance to the posed questions, and the responses varied significantly with each interaction. ChatGPT, in contrast to the aforementioned models, managed to provide responses that, while not always ideal, were generally adequate and maintained a level of consistency in delivering an output, even though the nature of the responses varied.

The agent debate phase was particularly influenced by the initial inconsistencies observed in Claude and the irrelevant outputs provided by Llama. As a result, the experimental debate was conducted exclusively using the ChatGPT model. This decision was predicated on the need for a more reliable dialogue interaction in this phase. During the debate, when discrepancies arose between the outcomes generated by the two agents at the end of the initial phase, a methodological approach was adopted wherein the agents progressed towards a unified conclusion by selecting the maximum from the set of priorities discussed. This approach facilitated a resolution that, while seeming synthetic, served to harmonize the divergent perspectives initially presented by the agents.

Inter-Rater Reliability in Prioritization Agent Performance Evaluation

The inter-rater reliability analysis of the prioritization agent's performance reveals mixed results across different evaluation criteria. While Clarity and Appropriateness demonstrate perfect agreement among reviewers with zero standard deviations and a Cohen's Kappa of 1, other criteria such as Relevance, Accuracy, and Coherence show moderate to significant variations in standard deviations and lower Kappa values. Notably, Consistency exhibits the most considerable discrepancies.

Table 4.5: Standard Deviations of Prioritization Agent Performance

Criterion	**GPT-4**	**Claude 3**	**Llama 3**	**All**
Relevance	0.47140452	0.47140452	0	0.41573971
Accuracy	0.47140452	0.47140452	0	0.47140452
Completeness	0	0	0.47140452	0.31426968
Clarity	0	0	0	0
Coherence	0	0.47140452	0.47140452	0.41573971
Appropriateness	0	0	0	0
Consistency	0.81649658	0	0	0.66666667

Table 4.6: Cohen's Kappa of Prioritization Agent Performance

Criterion	**GPT-4**	**Claude 3**	**Llama 3**	**All**
Relevance	0	0.4	1	0.625
Accuracy	0.4	0.5	1	0.65384615
Completeness	1	1	0	0.82
Clarity	1	1	1	1
Coherence	1	0.5	0	0.64
Appropriateness	1	1	1	1
Consistency	0	1	1	0.5

4.3.3 Internal Data Gathering Agent

Within the system, the Data Analyst agent is activated by the orchestration agent after human confirmation of the prioritization results. Its function is to verify the internal data availability for the prioritized suppliers and output a categorized list that identifies suppliers needing additional data via questionnaires and those for which sufficient internal data already exists.

Predefined Criteria of the Agent

Receiving two inputs: a prioritized list of components and supplier types, such as technology suppliers, and a list containing details about the suppliers (this is the internal knowledge that we have about the suppliers). The agent's task is to examine the supplier types and components marked as medium or high priority and identify which suppliers from the list align with these priorities. Additionally, the agent needs to determine which suppliers require the completion of questionnaires.

Input Data

The input data provided to the Agent includes two main components: a prioritized list of components and supplier types, and a detailed list of suppliers. This information encompasses details about their services, contact information, addresses, and the availability of a maturity assessment.

Here's an overview of the key elements of the input data:

1. Prioritized List:

   Components: High Priority Components, Medium Priority Components, Low Priority Components

   Supplier Types, such as Technology Suppliers

2. Detailed Supplier List:

   Each supplier entry includes contact details, an address (located in Amsterdam, Netherlands), a description of their specialized services, and the presence of a maturity assessment. The supplier descriptions elucidate the specific products and services offered, including card security features, mobile app development, authentication services, real-time payments, mortgage and loan products, customer service solutions, insurance, pension schemes, regulatory advice, and technology solutions.

Final Agent Prompt

You are the Internal Data Gathering Agent, responsible for categorizing suppliers based on a list of prioritized components and suppliers within the value chain. Your tasks include determining which specific suppliers belong to these prioritized components or suppliers and categorizing them into two categories: those requiring a questionnaire and those for whom the questionnaire can be skipped due to available maturity assessments. Follow the steps below to ensure accurate and efficient data gathering and categorization:

1. Receive Prioritized List: - Obtain the list of prioritized components and suppliers from the Orchestration Agent.

2. Identify Specific Suppliers: - Match the prioritized components and suppliers with specific suppliers from the supplier list.

3. Categorize Suppliers: - Categorize the previously identified suppliers with Medium and High Priority into two categories: 1. Questionnaire Required: Suppliers without available maturity assessments. 2. Questionnaire Not Required: Suppliers with available maturity assessments.

4. Check Maturity Assessment: - For each identified supplier, check if a maturity assessment is available in the internal data. - If a maturity assessment is available, categorize the supplier under "Questionnaire Not Required." - If a maturity assessment is not available, categorize the supplier under "Questionnaire Required."

5. Output Categorized List: - Generate a categorized list of suppliers, clearly indicating which suppliers fall under "Questionnaire Required" and which fall under "Questionnaire Not Required." - Provide a summary of the categorization process and any key findings.

Output Handling: - Ensure that the categorized list is clear, well-documented, and ready for further processing by the Orchestration Agent or other relevant agents. - Provide a summary of the categorization process and any key findings.

Prioritized List: <AGENT_INPUT>

Supplier List: <INTERNAL_DATA>

Complete one task at a time. At each step complete some parts of the tasks and summarize your findings at the end of your answer. If you receive the code: CONTINUE then continue with completing some parts of your task. Once you are done with all 5 steps then output the keyword: DONE

Evaluation System for the Internal Data Gathering Agent

1. **Relevance:**

   (a) **Grade 1:** The agent fails to focus on the prioritized components or suppliers, incorrectly identifying or ignoring critical data needs.

   (b) **Grade 2:** The agent generally identifies the correct suppliers and components but may include non-prioritized items or miss subtle distinctions between priority levels.

   (c) **Grade 3:** The agent accurately identifies and focuses exclusively on the prioritized components and suppliers, perfectly aligning with the set objectives.

2. **Accuracy:**

   (a) **Grade 1:** The agent incorrectly matches components and suppliers, leading to significant errors in the categorization process.

   (b) **Grade 2:** The agent mostly matches components and suppliers correctly but may make minor errors in identification that do not drastically affect the overall categorization.

   (c) **Grade 3:** The agent flawlessly matches components and suppliers with complete accuracy, ensuring that the categorization is entirely correct.

3. **Completeness:**

   (a) **Grade 1:** The agent provides incomplete or partial categorization of suppliers, missing critical elements required for subsequent processes.

   (b) **Grade 2:** The agent categorizes most suppliers accurately but may occasionally overlook some that require further data gathering or categorization.

   (c) **Grade 3:** The agent's categorization is comprehensive, including all relevant suppliers and ensuring that no required details are omitted.

4. **Clarity:**

   (a) **Grade 1:** The output is unclear or poorly organized, leading to confusion or misinterpretation in subsequent processing.

   (b) **Grade 2:** The output is mostly clear with occasional areas that might require additional clarification or better organization.

   (c) **Grade 3:** The agent's output is exceptionally clear, well-organized, and easy to follow, facilitating straightforward further processing.

5. **Coherence:**

   (a) **Grade 1:** The categorization logic is flawed or illogical, causing inconsistencies in how suppliers are grouped.

   (b) **Grade 2:** The agent's categorization is generally logical but may contain minor inconsistencies or irregularities.

   (c) **Grade 3:** The agent's process is highly coherent, with logical and consistent categorization that aligns well with the prioritization criteria.

6. **Appropriateness:**

(a) **Grade 1:** The tone or style of the agent's output is not suited for professional or operational use, potentially leading to misinterpretations.

(b) **Grade 2:** The agent's tone and style are largely appropriate, though there are occasional deviations that do not significantly impact its professional use.

(c) **Grade 3:** The agent maintains a professional tone and style throughout its output, enhancing usability and understanding.

7. **Consistency:**

(a) **Grade 1:** The agent's outputs vary significantly under identical conditions, leading to unreliable categorization.

(b) **Grade 2:** The agent displays some consistency in its outputs, with minor variations that do not majorly affect overall reliability.

(c) **Grade 3:** The agent consistently provides reliable and predictable categorizations, ensuring stability across multiple operations.

Outcome

Table 4.7: Qualitative Analysis of Internal Data Gathering Agent Performance

Criteria	**GPT-4**		**Claude 3**		**Llama 3**	
	Review 1	Review 2	Review 1	Review 2	Review 1	Review 2
Relevance	3	2.67	3	3	2	2
Accuracy	2	2	3	3	1	1
Completeness	2	2.3	3	2.67	1	1
Clarity	3	3	3	3	3	3
Coherence	2	2	3	3	1.67	2.3
Appropriateness	3	3	3	3	3	3
Consistency	3	2.67	3	3	3	2.67
Total Average	2.57	2.52	3	2.95	2.1	2.14

The Table 4.7 provided offers a comparative analysis of three AI models—GPT-4, Claude 3, and Llama 3—across several criteria, including relevance, accuracy, completeness, clarity, coherence, appropriateness, and consistency. The evaluation encompassed three distinct models, each tasked with the objective of categorizing suppliers into two groups: those necessitating the completion of questionnaires and those who could bypass this step and proceed directly to the transport data enrichment phase. All three models demonstrated proficiency in this classification task.

Llama 3 model, while capable of making the basic distinction between the two supplier categories, exhibited a notable limitation: it did not incorporate the prioritization list into its decision-making algorithm.

ChatGPT's performance was solely informed by the list of prioritized components failing to focus on the supplier types.

Claude, on the other hand, demonstrated a more nuanced understanding and application of the task requirements. It successfully utilized the list of prioritized components and suppliers to make informed selections. Claude's ability to integrate the list into its selection process suggests a higher level of sophistication and adaptability, positioning it as the most effective model among the three.

To assess the reliability of these models, a series of nine tests were conducted, with each model undergoing three tests. The consistency observed in the outcomes of these tests across all three models is indicative of their stability and reliability. Such consistency is crucial from a reliability standpoint, as it suggests that the models are likely to perform similarly under similar conditions in future applications.

Inter-Rater Reliability in Internal Data Gathering Agent Performance Evaluation

The analysis of the Internal Data Gathering Agent's performance shows varying agreement between reviewers across different evaluation criteria. Standard deviations indicate variability in certain areas, such as Relevance, Completeness, and Consistency, with values like 0.47140452 for GPT-4, suggesting moderate disagreement among reviewers. In contrast, criteria like Accuracy, Clarity, and Appropriateness show zero standard deviation and perfect Cohen's Kappa values of 1 across all models, indicating complete agreement. The criterion of Coherence displays significant variability in the Llama 3 model, with a standard deviation of 0.94280904 and a relatively lower Cohen's Kappa of 0.4, pointing to substantial differences in reviewer perceptions.

Table 4.8: Standard Deviations of Internal Data Gathering Agent Performance

Criterion	**GPT-4**	**Claude 3**	**Llama 3**	**All**
Relevance	0.47140452	0	0	0.31426968
Accuracy	0	0	0	0
Completeness	0.47140452	0.47140452	0	0.47140452
Clarity	0	0	0	0
Coherence	0	0	0.94280904	0.62853936
Appropriateness	0	0	0	0
Consistency	0.47140452	0	0.47140452	0.41573971

Table 4.9: Cohen's Kappa of Internal Data Gathering Agent Performance

Criterion	**GPT-4**	**Claude 3**	**Llama 3**	**All**
Relevance	0	1	1	0.75
Accuracy	1	1	1	1
Completeness	0.5	0	1	0.66666667
Clarity	1	1	1	1
Coherence	1	1	0.4	0.79545455
Appropriateness	1	1	1	1
Consistency	0	1	0	0

4.3.4 Questionnaire Generation Agent

The Questionnaire Generation Agent works with the Data Analyst Agent to collect supplier information and create questionnaires for data collection. It acts like a Cybersecurity Consultant, discussing and improving the questions to be asked.

Predefined Criteria

Based on the available data about a supplier generate a maturity assessment questionnaire that could be sent out to the suppliers to be filled out with the help of a chatbot.

Input Data

During the experiments conducted the input data provided describes the profile of AppTech Solutions, a mid-sized company specializing in developing mobile banking applications within the Financial Technology (FinTech) sector. The details include the company's size, industry specialization, headquarters location, and operational reach. Additionally, it outlines various compliance requirements with financial regulations and certifications, emphasizing the need to verify adherence to standards such as PCI DSS, GDPR, ISO 27001, and SOC 2, along with any applicable local and international regulations.

Final Agent Prompt

You are the Questionnaire Generation Agent, responsible for creating and customizing questionnaires for suppliers based on their specific attributes. Your tasks include analyzing the attributes of each supplier, adjusting the questions accordingly, and ensuring that the questionnaires are tailored to gather the most relevant information. Follow the steps below to generate and adjust the questionnaires:

1. Receive Supplier Information: - Obtain the list of suppliers from the Supplier Information Gathering Agent, along with their specific attributes (e.g., size, industry, geographical location, compliance requirements).

2. Analyze Supplier Attributes: - Review the attributes of each supplier to understand their unique characteristics and requirements.

3. Customize Questionnaire: - Adjust the questions in the questionnaire based on the supplier's attributes to ensure relevance and comprehensiveness. - Include additional questions or modify existing ones to address specific concerns related to the supplier's industry, size, or compliance needs.

4. Sample Questions: - Below are some sample questions. Customize these questions and create new ones as needed based on the supplier's attributes:

a. General Information: - What is the total number of employees in your organization? - Can you provide an overview of your business operations and primary products or services?

b. Compliance and Certifications: - Are you compliant with industry-specific regulations (e.g., GDPR, HIPAA)? Please provide details and documentation. - Do you have any relevant certifications (e.g., ISO 27001)? If so, please provide copies.

c. Security Practices: - What security measures do you have in place to protect sensitive data (e.g., encryption, access controls)? - Have you experienced any security incidents or data breaches in the past 24 months? If so, please describe the incident and the actions taken.

d. Risk Management: - How do you identify and manage risks within your organization? - Do you conduct regular risk assessments and vulnerability scans? Please provide details and recent reports.

e. Operational Practices: - How do you ensure the continuity of your business operations in case of a disruption (e.g., disaster recovery plans, business continuity plans)? - What measures do you take to ensure the quality and reliability of your products or services?

6. Outsourced operations and services - What critical security operations or services do you currently outsource, and to which providers? - How many data centers do you operate, and are any of these outsourced? Please specify locations and ownership details. - Do you engage any Cloud Service Providers (CSPs) for your operations? If so, list them and describe the scope of their services.

5. Finalize Questionnaire: - Review the customized questions for completeness and accuracy. - Ensure that the questions are clear and tailored to gather the necessary information from the supplier.

6. Output Questionnaire: - Generate the final questionnaire for each supplier.

Output Handling: - Ensure that the customized questionnaires are clear, well-documented, and tailored to each supplier's attributes. - Provide a summary of the customization process and any key considerations for each questionnaire.

Supplier information: <AGENT_INPUT>

Complete one task at a time. At each step complete some parts of the tasks and summarize your findings at the end of your answer. If you receive the code: CONTINUE then continue with completing some parts of your task. Once you are done with all 6 steps then output the keyword: DONE

Evaluation System for the Questionnaire Generation Agent

1. **Relevance:**
   (a) **Grade 1:** The agent generates questionnaires that are largely irrelevant to the supplier's profile, failing to address critical aspects of their operations or compliance needs.
   (b) **Grade 2:** The agent creates questionnaires that generally target relevant areas but may miss some specifics or include unnecessary questions.
   (c) **Grade 3:** The agent expertly tailors questionnaires to perfectly match the unique attributes and requirements of each supplier, ensuring all critical areas are covered.

2. **Accuracy:**
   (a) **Grade 1:** The questionnaire includes significant inaccuracies or outdated information that could lead to incorrect assessments of the supplier's capabilities or compliance.
   (b) **Grade 2:** The questionnaire is mostly accurate, with only minor errors that do not significantly impact the overall effectiveness of the data collection.
   (c) **Grade 3:** The questions are precise, well-researched, and accurately reflect the current standards and practices relevant to the supplier's industry and operational scope.

3. **Completeness:**
   (a) **Grade 1:** The questionnaire is incomplete, missing key questions that are essential for a thorough assessment of the supplier.
   (b) **Grade 2:** The questionnaire covers most areas necessary for assessment but may lack depth in some aspects or omit less obvious yet relevant queries.
   (c) **Grade 3:** The questionnaire is comprehensive, covering all necessary aspects of the supplier's operations, compliance, security practices, and risk management thoroughly.

4. **Clarity:**
   (a) **Grade 1:** The questions are ambiguous or confusing, potentially leading to misinterpretation or incomplete answers.
   (b) **Grade 2:** The questions are generally clear, though some may require rephrasing for better understanding or to elicit more precise responses.
   (c) **Grade 3:** Each question is formulated clearly and precisely, making it easy for suppliers to understand and respond accurately.

5. **Coherence:**
   (a) **Grade 1:** The sequence and grouping of questions are disorganized, leading to a disjointed or illogical flow that could confuse suppliers.

(b) **Grade 2:** The overall structure of the questionnaire is logical, but there could be improvements in how questions are grouped or sequenced for a smoother flow.

(c) **Grade 3:** The questionnaire is well-organized, with a logical flow that naturally progresses from general to specific, making it intuitive for suppliers to complete.

6. **Appropriateness:**

(a) **Grade 1:** The tone or style of the questionnaire is inappropriate, possibly too casual or too technical, which may affect the quality of responses.

(b) **Grade 2:** The tone is mostly appropriate, though some questions might benefit from adjustments to better suit the professional context of the supplier interaction.

(c) **Grade 3:** The questionnaire maintains a professional tone and is appropriately styled to engage suppliers effectively while ensuring that responses are informative and relevant.

7. **Consistency:**

(a) **Grade 1:** The questionnaires vary significantly in quality and relevance when applied to different suppliers under similar conditions.

(b) **Grade 2:** There is some variation in the consistency of the questionnaires, but these do not majorly affect the reliability of the information gathered.

(c) **Grade 3:** The agent consistently delivers high-quality, relevant questionnaires across different suppliers, ensuring reliable and uniform data collection.

Outcome

Table 4.10: Qualitative Analysis of Questionnaire Generation Agent Performance

Criteria	**GPT-4**		**Claude 3**		**Llama 3**	
	Review 1	Review 2	Review 1	Review 2	Review 1	Review 2
Relevance	3	3	1.67	2	1.3	1
Accuracy	2	2.3	1	1.3	1	1
Completeness	2	2	1	1	1	1
Clarity	3	3	3	3	3	3
Coherence	3	2	1	1	1	1
Appropriateness	3	3	3	3	3	3
Consistency	3	3	3	2.67	3	2
Total Average	2.71	2.61	1.95	2	1.9	1.71

The Table 4.10 provided offers a comparative analysis of three AI models—GPT-4, Claude 3, and Llama 3—across several criteria, including relevance, accuracy, completeness, clarity, coherence, appropriateness, and consistency. During the systematic execution of multiple iterations of the experiment, it was consistently observed that solely the ChatGPT model possessed the capability to generate coherent questions. A total of nine experiments were conducted, with each model being tested three times. Each instance of running the experimental procedures with ChatGPT yielded strikingly similar questions, illustrating its reliable performance in question generation under consistent conditions.

On the other hand, the Claude and Llama models demonstrated a distinct limitation in this area. Instead of producing questions, these models frequently offered only vague and general descriptions of what the task should entail, thereby failing to meet the experimental criteria for question generation.

Inter-Rater Reliability in Questionnaire Generation Agent Performance Evaluation

The inter-rater reliability analysis for the Questionnaire Generation Agent reveals both strengths and weaknesses in reviewer agreement across different criteria. Standard deviations suggest some discrepancies in criteria like Relevance, Accuracy, Coherence, and Consistency, with values indicating moderate to significant variation among the reviewers. For example, Coherence in GPT-4 shows a relatively high standard deviation of 0.81649658. In contrast, criteria such as Completeness, Clarity, and Appropriateness demonstrate perfect agreement, evidenced by zero standard deviations and a Cohen's Kappa of 1 across all models, indicating unanimous and consistent evaluations. Cohen's Kappa values for other criteria like Relevance, Accuracy, and Consistency show variability. Relevance and Consistency particularly exhibit lower kappa values in some models (e.g., Relevance in Llama 3 with a kappa of 0 and Consistency in Claude 3 with a kappa of 0), suggesting inconsistencies in how reviewers perceive and evaluate these aspects. The overall kappa for all models in these criteria (0.57142857 for Accuracy and 0.625 for Relevance) suggest a moderate to substantial agreement between the reviewers.

Table 4.11: Standard Deviations of Questionnaire Generation Agent Performance

Criterion	**GPT-4**	**Claude 3**	**Llama 3**	**All**
Relevance	0	0.47140452	0.47140452	0.47140452
Accuracy	0.47140452	0.47140452	0	0.41573971
Completeness	0	0	0	0
Clarity	0	0	0	0
Coherence	0.81649658	0	0	0.66666667
Appropriateness	0	0	0	0
Consistency	0	0.47140452	0.81649658	0.68493489

Table 4.12: Cohen's Kappa of Questionnaire Generation Agent Performance

Criterion	**GPT-4**	**Claude 3**	**Llama 3**	**All**
Relevance	1	0.5	0	0.625
Accuracy	0.5	0	1	0.57142857
Completeness	1	1	1	1
Clarity	1	1	1	1
Coherence	0	1	1	0.5

Table 4.12 continued from previous page

Criterion	GPT-4	Claude 3	Llama 3	All
Appropriateness	1	1	1	1
Consistency	1	0	0	0

4.3.5 Questionnaire Analyst Agent

The Data Analyst Agent in the system functions as a Task Planner that unifies and analyzes collected data for subsequent steps and identifies previously hidden suppliers. Based on its findings, it informs the Orchestration Agent to reinitiate the data collection process, incorporating these new suppliers.

Predefined Criteria

Upon receiving the completed questionnaire from the supplier, the agent's task is to analyze and consolidate the data for future steps and identify any suppliers not previously known in the system.

Input Data

The input data is a filled-out questionnaire. In the case of the experiments, the input data describes various operational aspects of AppTech Solutions, a company that develops mobile banking applications. A summarized overview:

1. General Information: This section offers a snapshot of the company's size, core focus, and methodologies used in software development, setting the stage for a deeper exploration of its operational specifics.

2. Compliance and Certifications: This part details the regulatory standards the company adheres to and the certifications it has achieved, highlighting areas for improvement in compliance management.

3. Security Practices: It describes the security measures currently implemented and identifies gaps in the security framework, particularly in access management and incident response.

4. Risk Management: This section covers the existing approaches to risk assessment and the need for a more robust framework, focusing on third-party integrations and ongoing vulnerability management.

5. Operational Practices: It discusses the company's strategies for ensuring business continuity and disaster recovery, along with the need for better quality assurance and operational adjustments to meet local regulatory requirements.

6. Security Operations and Outsourcing: This part outlines the outsourcing of certain cybersecurity functions and the operational specifics of data center management, providing insight into the company's resource allocation and external partnerships.

7. Cloud Services: The final section specifies the cloud platforms employed for hosting and development, illustrating the company's technological infrastructure and its use in supporting its business operations.

Final Agent Prompt

You are the Questionnaire Analyst Agent, responsible for analyzing the answers provided by suppliers to identify new suppliers in their supply chain and to score the risk associated with each response. Your tasks include reviewing the supplier's answers, identifying any additional suppliers mentioned, and assessing the risk level based on their responses. Follow the steps below to ensure thorough and accurate analysis and risk scoring:

1. Receive Supplier Responses: - Obtain the completed questionnaires from the supplier

2. Identify New Suppliers: - Carefully review the answers to identify any new suppliers or sub-suppliers mentioned by the supplier. - Extract relevant information about these new suppliers for further analysis and inclusion in the supply chain mapping.

3. Score Risk for Each Question: - Analyze each response to assess the risk level associated with it. - Use predefined risk assessment criteria to assign a risk score to each answer. Risk levels can be categorized as Low, Medium, High, or Critical.

4. Sample Risk Assessment Criteria: - Compliance and Certifications: - High risk: Non-compliance with essential regulations (e.g., PCI DSS, GDPR). - Medium risk: Partial compliance or missing some certifications. - Low risk: Full compliance with all relevant regulations and certifications.

- Security Practices: - High risk: No or inadequate security measures in place. - Medium risk: Basic security measures in place but lacking comprehensive controls. - Low risk: Robust security measures and practices implemented.

- Incident History: - High risk: Recent history of multiple or severe security incidents. - Medium risk: Few incidents with moderate impact. - Low risk: No recent security incidents or minor incidents with minimal impact.

- Risk Management: - High risk: Lack of regular risk assessments and vulnerability scans. - Medium risk: Occasional risk assessments but lacking thoroughness. - Low risk: Regular and comprehensive risk assessments and scans conducted.

- Operational Practices: - High risk: No disaster recovery or business continuity plans. - Medium risk: Basic plans in place but not regularly tested. - Low risk: Well-documented and regularly tested disaster recovery and business continuity plans.

5. Generate Risk Report: - Compile a report summarizing the identified new suppliers and the risk scores for each question. - Highlight any areas of high or critical risk that require immediate attention.

Output Handling: - Ensure that the risk assessment report is comprehensive and well-documented. - Provide a summary of the identified new suppliers and the risk scores for each question.

supplier response: <AGENT_INPUT>

Complete one task at a time. At each step complete some parts of the tasks and summarize your findings at the end of your answer. If you receive the code: CONTINUE then continue with completing some parts of your task. Once you are done with all 5 steps then output the keyword: DONE

Evaluation System for the Questionnaire Analyst Agent

1. **Relevance:**

   (a) **Grade 1:** The agent fails to focus on the crucial information needed to identify new suppliers or assess risks accurately, instead analyzing irrelevant details.

   (b) **Grade 2:** The agent generally identifies new suppliers and assesses risks but may miss some subtleties or overemphasize minor details.

   (c) **Grade 3:** The agent expertly identifies all new suppliers mentioned in the responses and accurately evaluates risks based on the responses, aligning perfectly with the objectives.

2. **Accuracy:**

   (a) **Grade 1:** The agent makes significant errors in risk assessment or fails to identify new suppliers mentioned in the questionnaire.

   (b) **Grade 2:** The agent is mostly accurate in its analysis but may have minor errors that do not drastically affect the overall risk assessment.

   (c) **Grade 3:** The agent's risk assessments and identification of new suppliers are completely accurate, reflecting a thorough understanding of the questionnaire content.

3. **Completeness:**

   (a) **Grade 1:** The analysis is incomplete, missing key risk areas or failing to identify critical new suppliers.
   (b) **Grade 2:** The agent covers most of the necessary areas for risk assessment and identifies many new suppliers, but some details might still be overlooked.
   (c) **Grade 3:** The agent's analysis is comprehensive, covering all aspects of the risk assessment and thoroughly identifying every new supplier.

4. **Clarity:**
   (a) **Grade 1:** The risk assessment reports are unclear or poorly organized, making it difficult to understand the risk levels or the implications of new suppliers.
   (b) **Grade 2:** The reports are mostly clear with some areas needing further clarification to enhance understanding or detail.
   (c) **Grade 3:** The risk assessment reports are exceptionally clear, well-organized, and easy to interpret, facilitating informed decision-making.

5. **Coherence:**
   (a) **Grade 1:** The analysis and reports are disjointed or illogical, with poor integration of findings that confuses the overall assessment.
   (b) **Grade 2:** The overall process and reports are logical, but minor inconsistencies may affect the flow or integration of the data.
   (c) **Grade 3:** The agent's analysis is logically sound and well-integrated, providing a coherent understanding of risks and supplier relationships.

6. **Appropriateness:**
   (a) **Grade 1:** The tone or style of the reports is inappropriate, which may mislead or fail to convey the seriousness of the risks.
   (b) **Grade 2:** The tone and style are mostly suitable, with only minor adjustments needed for better professionalism or impact.
   (c) **Grade 3:** The reports are consistently professional in tone and style, perfectly suited for the intended audience, enhancing the credibility and usefulness of the information.

7. **Consistency:**
   (a) **Grade 1:** The agent's outputs vary significantly under identical conditions, leading to unreliable risk assessments and supplier identification.
   (b) **Grade 2:** There is some variation in the agent's outputs, but these do not majorly affect the reliability of the information.
   (c) **Grade 3:** The agent delivers highly consistent and reliable assessments across different questionnaires, ensuring stability and predictability in its analyses.

Outcome

Table 4.13: Qualitative Analysis of Questionnaire Analyst Agent Performance

Criteria	**GPT-4**		**Claude 3**		**Llama 3**	
	Review 1	Review 2	Review 1	Review 2	Review 1	Review 2
Relevance	3	3	3	3	1.67	1.3
Accuracy	2.67	3	2.3	2	1	1
Completeness	3	3	3	3	1.3	1.3
Clarity	3	3	3	3	3	3
Coherence	3	2.67	3	3	1.3	2
Appropriateness	3	3	3	3	3	3
Consistency	3	2.67	3	3	1	1
Total Average	2.95	2.9	2.9	2.86	1.75	1.8

The Table 4.13 provided offers a comparative analysis of three AI models—GPT-4, Claude 3, and Llama 3—across several criteria, including relevance, accuracy, completeness, clarity, coherence, appropriateness, and consistency. Claude and ChatGPT exhibit proficiency in recognizing new suppliers and are capable of consistently generating comprehensive risk assessments, maintaining uniformity in their responses when prompted repeatedly. On the other hand, Llama, while capable of identifying new suppliers, produces risk assessments that are ad hoc and vary significantly with each instance. Occasionally, Llama's responses reach an acceptable level of quality, however, more frequently, the responses are suboptimal and yield information of minimal practical relevance. To evaluate the performance of these models thoroughly, a structured experimental setup was employed, comprising nine distinct tests, three for each model.

Inter-Rater Reliability in Questionnaire Analyst Agent Performance Evaluation

The inter-rater reliability analysis for the Questionnaire Analyst Agent reveals a mixed level of agreement among reviewers. Key areas such as Completeness, Clarity, and Appropriateness show perfect agreement with zero standard deviations and a Cohen's Kappa of 1, indicating consistent evaluations. However, criteria like Relevance, Accuracy, Coherence, and Consistency display some variability. Notably, Coherence has the lowest overall Kappa value of 0.34146341, pointing to differences in reviewer assessments. There is generally good agreement in areas like Relevance, Consistency, and Accuracy.

Table 4.14: Standard Deviations of Questionnaire Analyst Agent Performance

Criterion	**GPT-4**	**Claude 3**	**Llama 3**	**All**
Relevance	0	0	0.47140452	0.31426968
Accuracy	0.47140452	0.47140452	0	0.47140452
Completeness	0	0	0	0
Clarity	0	0	0	0
Coherence	0.47140452	0	0.47140452	0.56655772
Appropriateness	0	0	0	0

Table 4.14 continued from previous page

Criterion	GPT-4	Claude 3	Llama 3	All
Consistency	0.47140452	0	0	0.31426968

Table 4.15: Cohen's Kappa of Questionnaire Analyst Agent Performance

Criterion	GPT-4	Claude 3	Llama 3	All
Relevance	1	1	0.4	0.74285714
Accuracy	0	0.5	1	0.66666667
Completeness	1	1	1	1
Clarity	1	1	1	1
Coherence	0	1	0	0.34146341
Appropriateness	1	1	1	1
Consistency	0	1	1	0.78571429

4.3.6 Public Data Analyst Agent

The Data Analyst Agent in the system processes and refines public data, to identify and fill knowledge gaps sourced by the orchestration agent. It filters and analyzes this data. The output is an enhanced understanding of N-tier suppliers.

Predefined Criteria

Is able to process data to fill knowledge gaps in the risk assessment provided to it.

Input Data

The input data used during the experiments consisted of the risk assessment provided by the orchestration agent that was put together in the previous steps. Also example public data was given as input about the financial overview, market position, some key metrics like operating and profit margins, business segments, compliance certifications, key clients, recent developments, risk factors, key executives, strategic initiatives, supplier and partnership network, research and development, and ESG info.

Final Agent Prompt

You are the Public Data Analyst Agent, responsible for enhancing the analysis of suppliers by integrating and analyzing public data sources. Your tasks include reviewing supplier data from questionnaires and transport data analyses, and supplementing this information with insights from public data sources, such as S&P Capital IQ. Follow the steps below to ensure thorough and comprehensive analysis:

1. Receive Initial Data: - Obtain the supplier data from questionnaires and transport data analyses provided by the Orchestration Agent.

2. Integrate Public Data with Existing Analysis: - Cross-reference and integrate the public data with the information gathered from the questionnaires and transport data analysis. - Identify any discrepancies or additional insights that can enhance the overall supplier profile.

3. Sample Public Data Attributes to Analyze: - Financial Overview: - Revenue, net income, EBITDA, total assets, liabilities, and equity. - Market Position: - Market capitalization, share price, P/E ratio, sector, and subsector. - Compliance and Certifications: - PCI DSS, GDPR, ISO 27001, SOC 2, and other relevant certifications. - Key Clients and Partnerships: - Major clients and key partnerships that influence supplier reliability and stability. - Recent Developments: - Notable recent activities, such as product launches, market expansions, and investments. - Risk Factors: - Regulatory compliance risks, market competition, and technological changes. - Key Executives: - Profiles of top executives and their impact on company strategy.

4. Analyze and Score Risks: - Assess the risk level associated with each attribute based on predefined criteria. - Score the risk for each attribute as Low, Medium, High, or Critical.

5. Generate Enhanced Analysis Report: - Compile a comprehensive report that integrates the questionnaire responses, transport data analysis, and public data insights. - Highlight key findings, risk scores, and any new information that enhances the supplier profile.

Output Handling: - Ensure that the enhanced analysis report is comprehensive and well-documented. - Provide a summary of the integration process and any key findings.

Supplier Analysis: <AGENT_INPUT>

Public data: <S&P-Capital-IQ_INPUT>

Complete one task at a time. At each step complete some parts of the tasks and summarize your findings at the end of your answer. If you receive the code: CONTINUE then continue with completing some parts of your task. Once you are done with all 5 steps then output the keyword: DONE

Evaluation System for the Public Data Analyst Agent

1. **Relevance:**

   (a) **Grade 1:** The agent fails to integrate relevant public data or focuses on data that does not fill knowledge gaps in the risk assessment.

   (b) **Grade 2:** The agent generally uses relevant public data to enhance the supplier analysis but may include some irrelevant information or miss key data that could fill significant gaps.

   (c) **Grade 3:** The agent expertly utilizes public data to fill all significant knowledge gaps, providing a complete and enhanced understanding of N-tier suppliers.

2. **Accuracy:**

   (a) **Grade 1:** The integration of public data contains significant errors or misinterpretations that could mislead risk assessments.

   (b) **Grade 2:** The agent is mostly accurate in its data integration and analysis but may have minor inaccuracies that do not drastically impact the overall enhancement of the supplier profile.

   (c) **Grade 3:** The agent's data integration and analysis are completely accurate, reflecting a thorough understanding of public and internal data sources.

3. **Completeness:**

   (a) **Grade 1:** The analysis is incomplete, missing critical public data that is essential for a thorough assessment of the suppliers.

   (b) **Grade 2:** The agent covers most necessary public data but might occasionally overlook some details that could provide additional insights.

   (c) **Grade 3:** The agent's analysis is comprehensive, covering all necessary public data attributes and integrating them effectively with internal data.

4. **Clarity:**

   (a) **Grade 1:** The enhanced analysis report is unclear or poorly organized, making it difficult to understand the new insights or risk levels.

   (b) **Grade 2:** The report is mostly clear with some areas that could be better organized or more succinctly presented.

   (c) **Grade 3:** The enhanced analysis report is exceptionally clear, well-organized, and easy to interpret, facilitating informed decision-making.

5. **Coherence:**
   (a) **Grade 1:** The integration of public and internal data is disjointed or illogical, causing confusion in the overall assessment.
   (b) **Grade 2:** The overall integration process is logical but could be improved in terms of how data is correlated or presented.
   (c) **Grade 3:** The agent provides a highly coherent analysis, logically integrating public and internal data for a seamless understanding of supplier risks and profiles.
6. **Appropriateness:**
   (a) **Grade 1:** The tone or style of the analysis report is inappropriate, possibly too technical or casual, affecting its usefulness.
   (b) **Grade 2:** The tone and style are largely suitable, though some parts of the report might benefit from adjustments for better professionalism or readability.
   (c) **Grade 3:** The report maintains a professional tone and style throughout, perfectly suited for stakeholders and enhancing the credibility and utility of the information.
7. **Consistency:**
   (a) **Grade 1:** The agent's outputs vary significantly under identical conditions, leading to unreliable enhancements of the supplier profiles.
   (b) **Grade 2:** There is some variation in the agent's outputs, but these do not majorly affect the reliability of the enhancements.
   (c) **Grade 3:** The agent consistently delivers high-quality, reliable analyses across different data sets, ensuring stable and predictable enhancements of supplier information.

Outcome

Table 4.16: Qualitative Analysis of Public Data Analyst Agent Performance

Criteria	**GPT-4**		**Claude 3**		**Llama 3**	
	Review 1	Review 2	Review 1	Review 2	Review 1	Review 2
Relevance	3	3	2.67	3	1.3	1
Accuracy	2.67	2	2.3	2	1.3	1
Completeness	2.67	2.67	2.3	2	1	1
Clarity	3	3	3	3	3	3
Coherence	2	2.67	3	3	1.3	1.67
Appropriateness	3	3	3	3	3	3
Consistency	1	1	1	1	1	1
Total Average	2.48	2.48	2.47	2.43	1.7	1.67

The Table 4.16 provided offers a comparative analysis of three AI models—GPT-4, Claude 3, and Llama 3—across several criteria, including relevance, accuracy, completeness, clarity, coherence, appropriateness, and consistency. The models under review demonstrated a notable level of nondeterminism, with their outputs showing a significant degree of variability. Among the various models, ChatGPT emerged as the most effective, consistently delivering responses that were superior in quality, although it did exhibit occasional lapses in performance. In contrast, Claude managed to produce moderately acceptable answers in isolated instances, yet generally, its responses were largely inadequate and failed to meet a satisfactory standard. Llama was particularly notable for its consistent underperformance, not only failing to provide reliable answers but also producing outputs that were contradictory between different runs.

Inter-Rater Reliability in Public Data Analyst Agent Performance Evaluation

The inter-rater reliability analysis for the Public Data Analyst Agent shows a mixed pattern of agreement among the reviewers. Criteria like Clarity, Appropriateness, and Consistency display perfect agreement with zero standard deviation and a Cohen's Kappa of 1, indicating consistent evaluations by the reviewers across all models. However, Relevance, Accuracy, and Coherence exhibit variability in standard deviations and lower Cohen's Kappa scores. For example, Relevance and Accuracy have higher standard deviations around 0.47140452 and lower Kappa values (0.6 for Relevance and 0.36842105 for Accuracy), suggesting less agreement among reviewers. Coherence also presents notable inconsistencies with a Kappa of nearly zero for GPT-4, though it recovers slightly with the other models.

Table 4.17: Standard Deviations of Public Data Analyst Agent Performance

Criterion	**GPT-4**	**Claude 3**	**Llama 3**	**All**
Relevance	0	0.47140452	0.47140452	0.47140452
Accuracy	0.47140452	0.47140452	0.47140452	0.49690399
Completeness	0	0.47140452	0	0.31426968
Clarity	0	0	0	0
Coherence	0.47140452	0	0.47140452	0.47140452
Appropriateness	0	0	0	0
Consistency	0	0	0	0

Table 4.18: Cohen's Kappa of Public Data Analyst Agent Performance

Criterion	**GPT-4**	**Claude 3**	**Llama 3**	**All**
Relevance	1	0	0	0.6
Accuracy	0	0.5	0	0.36842105
Completeness	1	0.5	1	0.83333333
Clarity	1	1	1	1
Coherence	0	1	0.4	0.44897959
Appropriateness	1	1	1	1
Consistency	1	1	1	1

4.4 Technical Comparison and Performance Evaluation of the AI Models: GPT-4, Claude 3, and Llama 3

The performance of AI models like GPT-4, Claude 3, and Llama 3 is significantly shaped by their technical specifications, particularly the number of parameters, context window size, and maximum output tokens. Claude 3 operates with 137 billion parameters, Llama 3 with 70 billion, and GPT-4 with a substantially larger 1.7 trillion parameters. These differences directly impact each model's ability to process and analyze data. Additionally, the context window—200K for Claude 3, 8K for Llama 3, and 128K for GPT-4—further dictates the extent to which each model can handle large sequences of data effectively. The maximum output tokens also vary: Claude 3 allows for 8192 tokens, GPT-4 for 4096 tokens, and Llama 3 for 2048 tokens, influencing their capacity for generating extended responses. Given the black-box nature of these models, further details about their internal operations remain obscured, leaving these technical attributes as key determinants of performance. [6, 39]

GPT-4 emerged as the highest-performing model, demonstrating remarkable consistency and adaptability across a range of complex tasks. It excelled particularly in relevance, accuracy, and coherence, attributes likely bolstered by its extensive parameter count. GPT-4's superior performance in high-level reasoning and raw data processing is attributable to its large-scale architecture, which includes a significantly higher number of parameters than its counterparts.

When compared to GPT-4, Llama 3 generally underperformed, showing weaknesses in task relevance and consistency, which is likely a consequence of its considerably fewer parameters. For instance, in another comparative analysis by Ammar Ahmed, although both Llama 3 and GPT-4 were able to correctly identify and resolve a coding bug, GPT-4 provided a solution more closely aligned with the intended prompt in an image generation task. Additionally, in particularly complex tasks such as solving advanced mathematical problems, Llama 3 failed to deliver an accurate result. [6]

Claude 3, on the other hand, exhibited notable strengths, particularly in tasks requiring nuanced understanding and the integration of unstructured data, as demonstrated in internal data gathering tasks. This model outperformed GPT-4 in several specific areas, particularly those involving cognitive processing and complex reasoning. The model's advanced reasoning capabilities were highlighted in various benchmark tests, where Claude 3 Opus scored higher than GPT-4, particularly in graduate-level reasoning, coding tasks, and multilingual math. [39]

Concluding the above discussed, while GPT-4 excels in general adaptability and high-level reasoning, Claude 3 demonstrates particular strengths in tasks requiring sophisticated cognitive processing, coding, and reasoning. Llama 3, although faster in processing simpler tasks, underperforms in more complex scenarios, likely due to its smaller parameter set and limited context handling capacity.

Chapter 5

Discussion

In this chapter, we delve into the key findings of our study and explore the broader implications of deploying generative AI within multi-agent systems (MAS), particularly in the context of N-tier mapping. We examine the challenges faced during implementation, the limitations of the current study, and potential avenues for future research. Our discussion aims to provide a comprehensive understanding of the complexities and opportunities associated with this emerging technology.

5.1 Challenges

The deployment of generative AI in multi-agent systems introduces several significant challenges that must be addressed to harness its full potential. This section outlines the primary obstacles encountered, including issues related to explainability, determinism, hallucination, scalability, and learning and adaptation. Understanding these challenges is crucial for developing more robust and efficient MAS frameworks capable of performing complex tasks such as N-tier mapping.

5.1.1 Explainability

One of the critical challenges in deploying generative AI within multi-agent systems, particularly in the context of N-tier mapping, is the issue of explainability. Generative AI systems often operate as black boxes, making it difficult to interpret the decision-making processes of the agents involved. This lack of transparency can pose significant problems, especially when these systems are used in complex and dynamic environments where understanding the rationale behind decisions is crucial.

Explainability is essential for gaining trust and ensuring that the decisions made by AI agents can be understood and validated by human stakeholders. In the context of N-tier mapping, where agents are tasked with prioritizing components and suppliers, gathering and analyzing data, and enriching information with transport and public data, the ability to explain the outcomes and the processes leading to these outcomes is vital. Without sufficient explainability, it becomes challenging to diagnose errors, refine processes, and ensure compliance with regulatory and ethical standards.

5.1.2 Determinism

The issue of determinism is another significant challenge in the deployment of generative AI. Deterministic systems produce the same output given the same input, which is critical for predictability and reliability. However, generative AI systems, by their nature, can produce different results under the same conditions due to their probabilistic and adaptive models. This non-determinism introduces variability and unpredictability into the system.

On the other hand, while this variability can be seen as a disadvantage because it can lead to inconsistent outcomes, it also presents a unique advantage. The ability of generative AI systems to produce different outputs for the same input allows for multiple perspectives and solutions to a given problem.In the context of N-tier mapping, this means that agents can debate which results are better and why, allowing for various prioritizations and questionnaires to be compared and evaluated.

By analyzing these different outcomes, it is possible to refine processes and choose the best possible output. This iterative approach leads to continuous improvement and optimization of the system, making it more robust and effective over time. Moreover, the non-deterministic nature of these systems encourages the development of robust validation and selection mechanisms. By implementing rigorous evaluation criteria, suboptimal results can be filtered out, focusing on those that offer the highest quality and utility.

5.1.3 Hallucination

Hallucination is where models generate text that is not factually accurate or relevant to the given context. This issue is particularly pronounced in systems that utilize a single LLM-based agent but becomes exponentially more complicated in a system with multiple interacting agents.

Within a multi-agent framework, inaccurate outputs generated by a single agent can initiate a chain reaction of misinformation. This phenomenon is exacerbated by the interdependent nature of agents within these systems. Each agent typically relies on the inputs and outputs of other agents to make decisions, form strategies, and generate further responses. As a result, if one agent produces erroneous information, it can be inadvertently accepted as accurate by other agents. These agents then propagate the misinformation throughout the network, potentially leading to a widespread dissemination of false data.

The problem of hallucination in LLM-MA systems is further compounded by the complexity and dynamism of their operational environments. In contexts such as N-tier mapping, where agents are tasked with prioritizing components and suppliers, gathering and analyzing data, and enriching information with transport and public data, the accuracy and reliability of each agent's output are critical. A single hallucinated response can significantly undermine the integrity of the entire process, leading to incorrect prioritizations, flawed analyses, and misguided enrichment efforts.

Moreover, the collaborative nature of multi-agent systems means that agents often engage in debates to refine their outputs. Inaccurate information introduced into these debates can skew the discussions, resulting in suboptimal consensus and decision-making. The iterative nature of these debates, intended to enhance the quality of the final output, may instead amplify the impact of hallucinations if not properly managed.

Implementing cross-verification processes, where multiple agents independently assess the accuracy of generated information, can help mitigate the risk of misinformation propagation. Additionally, incorporating feedback loops from human overseers can provide an extra layer of scrutiny, ensuring that critical decisions are based on accurate and reliable data.

5.1.4 Scalability

LLM-MA systems are characterized by a substantial scalability challenge due to the sheer number of LLM-based agents required. Each agent, often developed on sophisticated language models like GPT-4, demands extensive computational power and memory. As the number of agents increases, so does the resource requirement, posing a significant hurdle in resource-limited scenarios.

5.1.5 Learning and Adaptation

Traditional multi-agent systems rely heavily on reinforcement learning from static, offline datasets. However, LLM-MA systems differ fundamentally in that they learn from dynamic, real-time feedback through continuous interactions with their environment or human users. This requirement for an interactive learning environment necessitates the design of stable and responsive systems capable of adapting to real-time inputs.

The current research practices in LLM-MA systems, including Memory and Self-Evolution techniques, are primarily focused on refining the behavior of individual agents based on feedback. While these methods are effective at improving the performance of single agents, they often fall short in leveraging the collective intelligence of the entire network of agents. This limitation arises because the strategies are typically agent-centric, not fully exploiting the synergistic benefits that coordinated multi-agent interactions can yield.

For example, memory modules allow agents to store and retrieve valuable past interactions, helping them make informed decisions based on historical data. However, these memories are often isolated within individual agents, preventing the sharing of useful experiences and knowledge across the network. Similarly, self-evolution techniques enable agents to dynamically adjust their strategies based on feedback, but this self-evolution is usually confined to the individual agent's perspective, missing opportunities for collaborative learning and adaptation.

To address these challenges, future research needs to focus on developing mechanisms that facilitate more effective collective learning and adaptation across the network. This could involve creating shared memory systems where agents can access and learn from each other's experiences, or developing frameworks that encourage collaborative problem-solving and strategy refinement.

5.2 Limitation

The rapidly evolving nature of generative AI, especially in the context of multi-agent systems, presents significant limitations for this study. The technology underpinning generative AI is still in its nascent stages, characterized by continual advancements and paradigm shifts. This research relies heavily on recent articles, many of which were published only a few months ago. Consequently, the findings and methodologies presented here may quickly become obsolete as new insights and technologies emerge.

One of the primary limitations is the potential for significant changes in the way we understand and implement generative AI in MAS. Given the swift pace of technological advancements, it is plausible that within a few months, the frameworks, models, and strategies discussed in this study may be rendered outdated or may require substantial revisions. This volatility necessitates continuous monitoring and adaptation of the research to keep pace with the latest developments.

Moreover, the novelty of generative AI in MAS introduces uncertainties in both theoretical and practical applications. As the technology matures, unforeseen challenges and opportunities will likely arise, necessitating further research to refine existing models and explore new paradigms. The current study provides a snapshot of the state-of-the-art but is inherently limited by its temporal context.

In this study, the implementation of MAS in N-tier mapping demonstrates the potential and challenges of utilizing generative AI within this framework. The orchestration agent, which oversees the entire process, interacts with various specialized agents to prioritize components and suppliers, gather and analyze data, and enrich findings with both transport and public data. Each agent's ability to adapt through self-evolution and memory modules highlights the dynamic capabilities of MAS. However, the effectiveness of these implementations is tied to the current state of technology, which is rapidly evolving.

The experimental design faced several constraints that may affect the generalizability and interpretation of the results. A primary limitation was the fixed temperature setting across the utilized generative AI models, namely ChatGPT-4, Claude 3, and LLama 3. The inability to adjust the temperature constrained the exploration of how varying levels of randomness could influence the models' performance.

Additionally, due to character limits imposed by these platforms, the experiments were conducted using shorter text inputs. This restriction meant that the study did not assess the models' effectiveness in processing longer inputs typical of real-world scenarios. This limitation poses questions about the scalability and adaptability of these models when faced with extended textual data, which remains untested within the confines of this research.

In light of these limitations, there is an urgent need for ongoing research to address the dynamic and rapidly changing nature of generative AI, particularly in its application to MAS and N-tier mapping. Future studies should focus on capturing the evolution of the technology and its applications in these contexts. This will help ensure that theoretical models and practical implementations remain relevant and effective.

Chapter 6

Conclusion

This thesis provides a comprehensive analysis of how N-tier mapping and generative AI can revolutionize supply chain management, specifically in enhancing cyber resilience and visibility. Through a detailed exploration of four main research questions, the thesis has highlighted the transformative potential of Large Language Models in traditional supply chain operations.

Firstly, by addressing *"What are the current strategies for cyber resilience specifically within cyber supply chain management?"* (Q1), the literature review in Chapter 2 established a foundational understanding of the field of cyber supply chain management. This review helped to contextualize current practices and emerging trends, setting a foundational understanding that enriched the analysis of generative AI integration and its potential to transform supply chain resilience and cyber security.

Following that, the examination of *"What is the current design and operational framework of N-tier mapping, and what areas could be enhanced to improve cyber resilience?"* (Q2) revealed that while existing strategies, including N-tier mapping, are effective, there are significant opportunities for improvement. The thesis identified enhancements necessary, as detailed in Section 3.1, for bolstering cyber resilience through advanced AI integration.

In response to *"What innovative approaches can GEN AI offer to enhance the N-tier process?"* (Q3), Section 3.2 introduced a multi-agent system that leverages various AI methodologies to advance operational frameworks. This innovative approach showcases how integrating LLMs enhances supply chain monitoring, providing sophisticated analysis and adaptive learning capabilities essential for improving cyber resilience.

The integration of LLMs within multi-agent systems, as discussed, represents an innovative approach. These AI models enhance the intelligence and efficacy of supply chain monitoring, providing sophisticated analysis, and adaptive learning capabilities that are critical for maintaining and enhancing cyber resilience.

Through the deployment of a Generative AI-powered chatbot and enhanced analysis of the value chain, the thesis further demonstrates how the multi-agent approach utilizing LLMs facilitates a more nuanced understanding of organizational maturity levels and critical focus areas. This enhanced capability addresses the knowledge gap by providing tailored recommendations and insights and improves visibility across the supply chain. Consequently, this leads to faster and more effective responses to disruptions, significantly enhancing cyber resilience.

Additionally, this thesis highlights a significant advancement in cyber resilience through the automation of the N-tier mapping process. Historically reliant on manual steps, the implementation of the multi-agent system now automates this process, reducing human error and time needed, and increasing efficiency.

Addressing the final research question *"What are the potential challenges and limitations of integrating GEN AI into this system?"* (Q4), in Chapter 4 experiments are conducted with AI models such as GPT-4, Claude 3, and Llama 3 within a simulated N-tier mapping scenario revealed distinct strengths and limitations of each model. The findings indicate that while ChatGPT (GPT-4) generally exhibits consistent and adaptable performance across a range of tasks, the variability in the output of other models highlights the challenges of determinism and hallucination. These experiments provide critical insights into the capability of LLMs to handle complex, multi-layered tasks and their role in enhancing the functionality of MAS in real-world scenarios.

The thesis identified key challenges such as explainability, determinism, hallucination, scalability, and the need for ongoing adaptation in Chapter 5. These challenges underline the complexities involved in deploying AI in real-world scenarios and emphasize the necessity for continued research and development.

Concluding the above discussed, this thesis highlights the transformative potential of integrating advanced AI into supply chain management and maps out a future where continuous technological advancements and methodological innovations could address current limitations. The potential for AI to adapt and evolve with the changing dynamics of global supply chains offers a promising horizon for research and application.

Bibliography

[1] Reimagining engineering manufacturing and supply chain report 2024. Technical report, 2024.

[2] Accenture. Securing the Supply Chain, 2020. Accessed: 15/03/22024. URL: https://www.accenture.com/content/dam/accenture/final/a-com-migration/r3-3/pdf/pdf-134/accenture-securing-the-supply-chain.pdf#zoom=40.

[3] Accenture. How visibility boosts supply chain resilience. https://www.accenture.com/us-en/blogs/high-tech/how-visibility-boosts-supply-chain-resilience, 2022.

[4] Accenture. Various internal documents on cyber supply chain risk management. Internal documents, 2024. Confidential: not publicly accessible.

[5] Giuseppe Aceto, Valerio Persico, and Antonio Pescapé. The role of information and communication technologies in healthcare: taxonomies, perspectives, and challenges. *Journal of Network and Computer Applications*, 107:125–154, 2018. URL: https://www.sciencedirect.com/science/article/pii/S1084804518300456, doi:10.1016/j.jnca.2018.02.008.

[6] Ammar Ahmed. Meta ai llama 3 vs. chatgpt 4: A comparative analysis of ai assistant capabilities. *Article*, May 2024. Accessed: 2024-07-20.

[7] Fatima Alwahedi, Alyazia Aldhaheri, Mohamed Amine Ferrag, Ammar Battah, and Norbert Tihanyi. Machine learning techniques for iot security: Current research and future vision with generative ai and large language models. *Internet of Things and Cyber-Physical Systems*, 2024.

[8] Asad Arfeen, Saad Ahmed, Muhammad Asim Khan, and Syed Faraz Ali Jafri. Endpoint detection response: A malware identification solution. In *2021 International Conference on Cyber Warfare and Security (ICCWS)*, pages 1–8, 2021. doi:10.1109/ICCWS53234.2021.9703010.

[9] Sandeep Bhatt, Pratyusa K. Manadhata, and Loai Zomlot. The operational role of security information and event management systems. *IEEE Security Privacy*, 12(5):35–41, 2014. doi:10.1109/MSP.2014.103.

[10] Jon M. Boyens, Angela Smith, Nadya Bartol, Kris Winkler, Alexander Holbrook, and Matthew Fallon. Cybersecurity supply chain risk management for systems and organizations, 2022-05-05 04:05:00 2022. URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=934690, doi:10.6028/NIST.SP.800-161r1.

[11] Thomas M. Chen and Saeed Abu-Nimeh. Lessons from stuxnet. *Computer*, 44(4):91–93, 2011. doi:10.1109/MC.2011.115.

[12] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. volume 13-17-August-2016, page 785 – 794, 2016. Cited by: 23092; All Open Access, Bronze Open Access, Green Open Access. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84984950690&doi=10.1145%2f2939672.2939785&partnerID=40&md5=73effb39b7405c4cd7e8da0e496f6de7, doi:10.1145/2939672.2939785.

[13] Konstantinos Christidis and Michael Devetsikiotis. Blockchains and smart contracts for the internet of things. *IEEE Access*, 4:2292–2303, 2016. doi:10.1109/ACCESS.2016.2566339.

[14] J.W. Creswell. *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*. Pearson, 2012. URL: https://books.google.nl/books?id=4PywcQAACAAJ.

[15] Ajay Das, Simone Gottlieb, and Dmitry Ivanov. *Managing Disruptions and the Ripple Effect in Digital Supply Chains: Empirical Case Studies*, pages 261–285. Springer International Publishing, Cham, 2019. doi:10.1007/978-3-030-14302-2_13.

[16] Derui Ding, Qing-Long Han, Zidong Wang, and Xiaohua Ge. A survey on model-based distributed control and filtering for industrial cyber-physical systems. *IEEE Transactions on Industrial Informatics*, 15(5):2483–2499, 2019. doi:10.1109/TII.2019.2905295.

[17] Derui Ding, Qing-Long Han, Yang Xiang, Xiaohua Ge, and Xian-Ming Zhang. A survey on security control and attack detection for industrial cyber-physical systems. *Neurocomputing*, 275:1674–1683, 2018. URL: https://www.sciencedirect.com/science/article/pii/S0925231217316351, doi:10.1016/j.neucom.2017.10.009.

[18] Alexandre Dolgui Dmitry Ivanov and Boris Sokolov. The impact of digital technology and industry 4.0 on the ripple effect and supply chain risk analytics. *International Journal of Production Research*, 57(3):829–846, 2019. arXiv:https://doi.org/10.1080/00207543.2018.1488086, doi:10.1080/00207543.2018.1488086.

[19] Naveen Donthu, Satish Kumar, Debmalya Mukherjee, Nitesh Pandey, and Weng Marc Lim. How to conduct a bibliometric analysis: An overview and guidelines. *Journal of business research*, 133:285–296, 2021.

[20] Min Du, Feifei Li, Guineng Zheng, and Vivek Srikumar. Deeplog: Anomaly detection and diagnosis from system logs through deep learning. page 1285 – 1298, 2017. Cited by: 936. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85041449815&doi=10.1145%2f3133956.3134015&partnerID=40&md5=39da60f53442ae3ae47f2e7ccb95075e, doi:10.1145/3133956.3134015.

[21] Niloofar Etemadi, Pieter Van Gelder, and Fernanda Strozzi. An ism modeling of barriers for blockchain/distributed ledger technology adoption in supply chains towards cybersecurity. *Sustainability*, 13(9), 2021. URL: https://www.mdpi.com/2071-1050/13/9/4672, doi:10.3390/su13094672.

[22] Huanhuan Feng, Xiang Wang, Yanqing Duan, Jian Zhang, and Xiaoshuan Zhang. Applying blockchain technology to improve agri-food traceability: A review of development methods, benefits and challenges. *Journal of Cleaner Production*, 260:121031, 2020. URL: https://www.sciencedirect.com/science/article/pii/S0959652620310787, doi:10.1016/j.jclepro.2020.121031.

[23] Mohamed Amine Ferrag, Mthandazo Ndhlovu, Norbert Tihanyi, Lucas C Cordeiro, Merouane Debbah, Thierry Lestable, and Narinderjit Singh Thandi. Revolutionizing cyber threat detection with large language models: A privacy-preserving bert-based lightweight model for iot/iiot devices. *IEEE Access*, 2024.

[24] Abhijeet Ghadge, Maximilian Weiß, Nigel D Caldwell, and Richard Wilding. Managing cyber risk in supply chains: A review and research agenda. *Supply Chain Management: An International Journal*, 25(2):223–240, 2019.

[25] Jairo Giraldo, David Urbina, Alvaro Cardenas, Junia Valente, Mustafa Faisal, Justin Ruths, Nils Ole Tippenhauer, Henrik Sandberg, and Richard Candell. A survey of physics-based attack detection in cyber-physical systems. *ACM Computing Surveys*, 51(4), 2018. Cited by: 289; All Open Access, Bronze Open Access, Green Open Access. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85053229533&doi=10.1145%2f3203245&partnerID=40&md5=8f20f4ae8b446cd53d5751dc1141ff72, doi:10.1145/3203245.

[26] Jairo Giraldo, David Urbina, Alvaro Cardenas, Junia Valente, Mustafa Faisal, Justin Ruths, Nils Ole Tippenhauer, Henrik Sandberg, and Richard Candell. A survey of physics-based attack detection in cyber-physical systems. *ACM Comput. Surv.*, 51(4), jul 2018. doi:10.1145/3203245.

[27] Joshua Glasser and Brian Lindauer. Bridging the gap: A pragmatic approach to generating insider threat data. page 98 – 104, 2013. Cited by: 191; All Open Access, Bronze Open Access. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84882742580&doi=10.1109%2fSPW.2013.37&partnerID=40&md5=7f1897dc2c799031a3a8e93bec05d858, doi:10.1109/SPW.2013.37.

[28] Kannan Govindan, Hassan Mina, and Behrouz Alavi. A decision support system for demand management in healthcare supply chains considering the epidemic outbreaks: A case study of coronavirus disease 2019 (covid-19). *Transportation Research Part E: Logistics and Transportation Review*, 138:101967, 2020. URL: https://www.sciencedirect.com/science/article/pii/S1366554520306189, doi:10.1016/j.tre.2020.101967.

[29] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

[30] Mohammad Sayad Haghighi, Faezeh Farivar, and Alireza Jolfaei. A machine-learning-based approach to build zero-false-positive ipss for industrial iot and cps with a case study on power grids security. *IEEE Transactions on Industry Applications*, 60(1):920–928, 2024. doi:10.1109/TIA.2020.3011397.

[31] Nils-Ole Hohenstein, Edda Feisel, Evi Hartmann, and Larry Giunipero. Research on the phenomenon of supply chain resilience: a systematic review and paths for further investigation. *International journal of physical distribution & logistics management*, 45(1/2):90–117, 2015.

[32] Dmitry Ivanov. Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (covid-19/sars-cov-2) case. *Transportation Research Part E: Logistics and Transportation Review*, 136:101922, 2020.

[33] Dmitry Ivanov, Alexandre Dolgui, Ajay Das, and Boris Sokolov. *Digital Supply Chain Twins: Managing the Ripple Effect, Resilience, and Disruption Risks by Data-Driven Optimization, Simulation, and Visibility*, pages 309–332. Springer International Publishing, Cham, 2019. doi:10.1007/978-3-030-14302-2_15.

[34] Dmitry Ivanov, Alexandre Dolgui, Boris Sokolov, Frank Werner, and Marina Ivanova. A dynamic model and an algorithm for short-term supply chain scheduling in the smart factory industry 4.0. *International Journal of Production Research*, 54(2):386 – 402, 2016. Cited by: 411; All Open Access, Green Open Access. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84955677643&doi=10.1080%2f00207543.2014.999958&partnerID=40&md5=e3b74d3c0e08fcde346e8cfbbe09a826, doi:10.1080/00207543.2014.999958.

[35] Aram Kim, Junhyoung Oh, Jinho Ryu, and Kyungho Lee. A review of insider threat detection approaches with iot perspective. *IEEE Access*, 8:78847–78867, 2020. doi:10.1109/ACCESS.2020.2990195.

[36] Xiangyuan Lan, Wei Zhang, Shengping Zhang, Deepak Kumar Jain, and Huiyu Zhou. Robust multi-modality anchor graph-based label prediction for rgb-infrared tracking. *IEEE Transactions on Industrial Informatics*, pages 1–1, 2019. doi:10.1109/TII.2019.2947293.

[37] Mamoona Majid, Shaista Habib, Abdul Rehman Javed, Muhammad Rizwan, Gautam Srivastava, Thippa Reddy Gadekallu, and Jerry Chun-Wei Lin. Applications of wireless sensor networks and internet of things frameworks in the industry revolution 4.0: A systematic literature review. *Sensors*, 22(6), 2022. URL: https://www.mdpi.com/1424-8220/22/6/2087, doi:10.3390/s22062087.

[38] Imran Makhdoom, Mehran Abolhasan, Justin Lipman, Ren Ping Liu, and Wei Ni. Anatomy of threats to the internet of things. *IEEE Communications Surveys Tutorials*, 21(2):1636–1675, 2019. doi:10.1109/COMST.2018.2874978.

[39] Chris Mann and Anita Kirkovska. Claude 3 opus vs gpt-4: Task specific analysis. *Article*, April 8 2024. Accessed: 2024-07-20.

[40] Stefan Mihai, Mahnoor Yaqoob, Dang V. Hung, William Davis, Praveer Towakel, Mohsin Raza, Mehmet Karamanoglu, Balbir Barn, Dattaprasad Shetve, Raja V. Prasad, Hrishikesh Venkataraman, Ramona Trestian, and Huan X. Nguyen. Digital twins: A survey on enabling technologies, challenges, trends and future prospects. *IEEE Communications Surveys Tutorials*, 24(4):2255–2291, 2022. doi:10.1109/COMST.2022.3208773.

[41] Saikat Mondal, Kanishka P. Wijewardena, Saranraj Karuppuswami, Nitya Kriti, Deepak Kumar, and Premjeet Chahal. Blockchain inspired rfid-based information architecture for food supply chain. *IEEE Internet of Things Journal*, 6(3):5803–5813, 2019. doi:10.1109/JIOT.2019.2907658.

[42] David Mudzingwa and Rajeev Agrawal. A study of methodologies used in intrusion detection and prevention systems (idps). In *2012 Proceedings of IEEE Southeastcon*, pages 1–6, 2012. doi:10.1109/SECon.2012.6197080.

[43] Khan Muhammad, Rafik Hamza, Jamil Ahmad, Jaime Lloret, Haoxiang Wang, and Sung Wook Baik. Secure surveillance framework for iot systems using probabilistic

image encryption. *IEEE Transactions on Industrial Informatics*, 14(8):3679–3689, 2018. doi:10.1109/TII.2018.2791944.

[44] Arwa Mukhtar, Awanis Romli, and Noorhuzaimi Karimah Mohd. Blockchain network model to improve supply chain visibility based on smart contract. *International Journal of Advanced Computer Science and Applications*, 11(10), 2020.

[45] Rida Nasir, Mehreen Afzal, Rabia Latif, and Waseem Iqbal. Behavioral based insider threat detection using deep learning. *IEEE Access*, 9:143266 – 143274, 2021. Cited by: 25; All Open Access, Gold Open Access. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85118997499&doi=10.1109%2fACCESS.2021.3118297&partnerID=40&md5=73e09412156533dc28cd09537008fecf, doi:10.1109/ACCESS.2021.3118297.

[46] Andrew Ng. Agentic design patterns part 1. https://www.deeplearning.ai/the-batch/how-agents-can-improve-llm-performance/, March 2024. Accessed: 2024-04-20.

[47] Andrew Ng. Agentic design patterns part 2, reflection. https://www.deeplearning.ai/the-batch/agentic-design-patterns-part-2-reflection/, March 2024. Accessed: 2024-04-20.

[48] Andrew Ng. Agentic design patterns part 3, tool use. https://www.deeplearning.ai/the-batch/agentic-design-patterns-part-3-tool-use/, April 2024. Accessed: 2024-04-20.

[49] Andrew Ng. Agentic design patterns part 4, planning. https://www.deeplearning.ai/the-batch/agentic-design-patterns-part-4-planning/, April 2024. Accessed: 2024-04-20.

[50] Andrew Ng. Agentic design patterns part 5, multi-agent collaboration. https://www.deeplearning.ai/the-batch/agentic-design-patterns-part-5-multi-agent-collaboration/, April 2024. Accessed: 2024-04-20.

[51] Ethan Nikookar and Yoshio Yanadori. Preparing supply chain for the next disruption beyond covid-19: managerial antecedents of supply chain resilience. *International journal of operations & production management*, 42(1):59–90, 2022.

[52] Mehrdokht Pournader, Andrew Kach, and Srinivas Talluri. A review of the existing and emerging topics in the supply chain risk management literature. *Decision sciences*, 51(4):867–919, 2020.

[53] Petar Radanliev, David De Roure, Kevin Page, Jason R. C. Nurse, Rafael Mantilla Montalvo, Omar Santos, La'Treall Maddox, and Pete Burnap. Cyber risk at the edge: current and future trends on cyber risk analytics and artificial intelligence in the industrial internet of things and industry 4.0 supply chains. *Cybersecurity*, 3(1), 2020. Cited by: 63; All Open Access, Gold Open Access, Green Open Access. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089683692&doi=10.1186%2fs42400-020-00052-8&partnerID=40&md5=3e96294c154e73fc2ca83664b8207a53, doi:10.1186/s42400-020-00052-8.

[54] Chris Rickman and Alex Chao. Microagents: Exploring agentic architecture with microservices. https://devblogs.microsoft.com/semantic-kernel/microagents-exploring-agentic-architecture-with-microservices/, January 2024. Accessed: 2024-04-06.

[55] Deepak Sharma, Ruchi Mittal, Ravi Sekhar, Pritesh Shah, and Matthias Renz. A bibliometric analysis of cyber security and cyber forensics research. *Results in Control and Optimization*, 10:100204, 2023. URL: https://www.sciencedirect.com/science/article/pii/S2666720723000061, doi:10.1016/j.rico.2023.100204.

[56] Karamveer Singh Sidhu. Top 5 famous software supply chain attacks in 2023: Explore the critical nature of supply chain cyber attacks and learn how to fortify your defenses against this growing threat in 2023. *Unknown Journal*, November 24 2023.

[57] Alexander Sniffin. Three ai design patterns of autonomous agents. https://alexsniffin.medium.com/three-ai-design-patterns-of-autonomous-agents-8372b9402f7c, March 2024. Accessed: 2024-04-06.

[58] Alexander Spieske and Hendrik Birkel. Improving supply chain resilience through industry 4.0: A systematic literature review under the impressions of the covid-19 pandemic. *Computers Industrial Engineering*, 158:107452, 2021. URL: https://www.sciencedirect.com/science/article/pii/S0360835221003569, doi:10.1016/j.cie.2021.107452.

[59] Caroline Swift, V Daniel R Guide Jr, and Suresh Muthulingam. Does supply chain visibility affect operating performance? evidence from conflict minerals disclosures. *Journal of Operations Management*, 65(5):406–429, 2019.

[60] Sheetal Thakare, Anshuman Pund, and M. A. Pund. Network traffic analysis, importance, techniques: A review. In *2018 3rd International Conference on Communication and Electronics Systems (ICCES)*, pages 376–381, 2018. doi:10.1109/CESYS.2018.8723955.

[61] Aaron Tuor, Samuel Kaplan, Brian Hutchinson, Nicole Nichols, and Sean Robinson. Deep learning for unsupervised insider threat detection in structured cybersecurity data streams. volume WS-17-01 - WS-17-15, page 224 – 234, 2017. Cited by: 135. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85046103696&partnerID=40&md5=3d46268ca6592e08c93af802b75f57e3.

[62] Yoeri van Bruchem. Qualitative evaluation of llm responses. *Ordina Data*, 4 2024. Accessed: 2024-05-12.

[63] Viswanath Venkatesh, Susan A Brown, and Hillol Bala. Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems. *MIS quarterly*, pages 21–54, 2013.

[64] Zhi Qiang Wang and Abdulmotaleb El Saddik. Dtitd: An intelligent insider threat detection framework based on digital twin and self-attention based deep learning models. *IEEE Access*, 2023.

[65] Pascal Wichmann, Alexandra Brintrup, Simon Baker, Philip Woodall, and Duncan McFarlane. Extracting supply chain maps from news articles using deep neural networks. *International Journal of Production Research*, 58(17):5320–5336, 2020.

[66] Hansong Xu, Jun Wu, Qianqian Pan, Xinping Guan, and Mohsen Guizani. A survey on digital twin for industrial internet of things: Applications, technologies and tools. *IEEE Communications Surveys Tutorials*, 25(4):2569–2598, 2023. doi:10.1109/COMST.2023.3297395.

[67] Fangfang Yuan, Yanan Cao, Yanmin Shang, Yanbing Liu, Jianlong Tan, and Binxing Fang. Insider threat detection with deep neural network. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10860 LNCS:43 – 54, 2018. Cited by: 83. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85048985680&doi=10.1007%2f978-3-319-93698-7_4&partnerID=40&md5=9e6de341c06a2ad2e1ee705597f064e9, doi:10.1007/978-3-319-93698-7_4.

[68] Junlong Zhou, Liying Li, Ahmadreza Vajdi, Xiumin Zhou, and Zebin Wu. Temperature-constrained reliability optimization of industrial cyber-physical systems using machine learning and feedback control. *IEEE Transactions on Automation Science and Engineering*, 20(1):20–31, 2023. doi:10.1109/TASE.2021.3062408.