Service Oriented Computing and Applications
https://doi.org/10.1007/s11761-025-00474-7

ORIGINAL RESEARCH PAPER

Benchmarking large language models for supply chain risk identification: an extended evaluation within the LARD-SC framework

Ming Zhao[1] · Omar Hussain[1] · Yu Zhang[1] · Morteza Saberi[2] · Abderrahmane Leshob[3]

Received: 23 January 2025 / Revised: 11 April 2025 / Accepted: 19 September 2025

Abstract
Operational resilience in modern global supply chains depends on timely and accurate identification of emerging risks. While daily news has become a primary source for such insights, the sheer volume and unstructured nature of these data pose significant analytical challenges, requiring advanced tools to extract relevant and actionable information. This paper introduces an extended evaluation of the LARD-SC framework, a service-oriented architecture for supply chain risk management, by benchmarking five diverse variants of the large language model (LLM) in their capacity to detect, classify, and interpret risks. Drawing on a curated set of 120 real-world news articles on Apple's Tier 1 suppliers, we adopt a standardized, prompt-based assessment to compare GPT-3.5 turbo, GPT-4o, GPT-4o mini, Claude 3.5 Sonnet, and Claude 3.5 Haiku. Using expert-reviewed metrics, namely the Risk Validation Rate (RVR), Potential Risk Rate (PRR), and False Identification Rate (FIR), we derive a comprehensive Relative Performance Index (RPI) for comparison. Our analysis confirms that advanced GPT-4o variants produce the most consistent accurate risk identifications, achieving higher proportions of validated outcomes while minimizing false positives. Through these results, we highlight the significant promise of LLM-driven analytics for early risk detection in complex supply chains, along with practical considerations such as the influence of prompt engineering, interpretability demands, and the impact of data availability. The findings offer a blueprint for organizations seeking to improve resilience by systematically harnessing the capabilities of LLM within service-oriented risk management ecosystems.

Keywords Service-oriented computing · Supply chain risk management · Large language models · Predictive analytics

1 Key terms

1.1 Risk manager (RM)

A *risk manager (RM)* is the individual or team responsible for identifying, assessing, and mitigating disruptions or uncertainties that may compromise the focal company's operational continuity and strategic goals. RMs leverage data-driven insights such as those generated by LLMs, to make informed decisions regarding supplier vulnerabilities, contingency planning, and crisis response.

1.2 Supply chain network (SCN)

A *supply chain network (SCN)* is the comprehensive, interconnected system comprising organizations, resources, infrastructures, and information flows that collectively facilitate the production and delivery of goods and services from the initial procurement of raw materials to the ultimate distribution to end consumers. This network encompasses multiple

✉ Ming Zhao
ming.zhao@unsw.edu.au

Omar Hussain
o.hussain@unsw.edu.au

Yu Zhang
m.yuzhang@unsw.edu.au

Morteza Saberi
morteza.saberi@uts.edu.au

Abderrahmane Leshob
leshob.abderrahmane@uqam.ca

[1] School of Business, UNSW Canberra, Canberra, Australia

[2] School of Computer Science, University of Technology Sydney, Sydney, Australia

[3] School of Management, University of Quebec at Montreal, Montreal, Canada

Published online: 29 September 2025

tiers of suppliers, including, but not limited to, tier-1 suppliers as well as manufacturers, logistics providers, distributors, and retail channels, each representing critical nodes whose interdependencies significantly influence overall system performance.

1.3 Supply chain risk management (SCRM)

Supply chain risk management (SCRM) encompasses the systematic processes and strategies aimed at identifying, evaluating, and mitigating risks across the interconnected stages of supply, manufacturing, distribution, and delivery. SCRM seeks to maintain operational resilience and protect organizational objectives from disruptions arising from internal and external factors (e.g., supplier failures and geopolitical events).

1.4 Large language model (LLM)

A *large language model (LLM)* is an artificial intelligence system, typically built with deep neural network architectures, trained on vast textual corpora. LLMs, such as GPT-4o, are adept at performing various language-related tasks, including text generation, summarization, sentiment analysis, and information extraction.

1.5 Large language model-based approach for risk identification (LLM-RI)

Large language model-based approach for Risk Identification (LLM-RI) is a sub-framework within LARD-SC that uses LLMs to automatically detect potential disruptions or vulnerabilities from unstructured textual data (e.g., online news articles). Through specialized prompt engineering, it assesses the likelihood and impact of each identified risk event, enabling real-time alerts to risk managers.

1.6 Large language model-based approach for risk classification (LLM-RC)

Large Language Model-based approach for Risk Classification (LLM-RC) is a sub-framework within LARD-SC responsible for systematically categorizing the risks identified by LLM-RI. Applying semantic embeddings and leveraging a standardized taxonomy (the Cambridge Taxonomy of Business Risks) translates free-text risk descriptions into consistent, interpretable labels that facilitate prioritization and decision making.

1.7 Data collection and visualization for risk analysis (DCV-RA)

Data Collection and Visualization for Risk Analysis (DCV-RA) is the sub-framework within LARD-SC that orchestrates the automated gathering of supplier-related information, integrates the results of LLM-RI and LLM-RC into a centralized database, and offers a visual interface, powered by graph databases like Neo4j to risk managers. DCV-RA provides an interactive environment for exploring supplier relationships, identified risks, and their interconnections.

1.8 Cambridge taxonomy of business risks (CTBR)

The *Cambridge Taxonomy of Business Risks (CTBR)* is a hierarchical classification system used to organize a diverse range of business disruptions, including financial, geopolitical, technological, environmental, social, and governance risks. Defined by the Cambridge Centre for Risk Studies [1], the CTBR maps textual descriptions of disruptions to a stable reference system. In doing so, it promotes consistency and enhances the comparability of risk information across different supply chain contexts.

2 Introduction

In an era marked by unprecedented global interconnectivity and rapidly evolving disruptions, effective supply chain risk management (SCRM) has become vital for maintaining operational resilience and securing competitive advantages. Supply chains today face a multifaceted landscape of uncertainties, from supplier failures and geopolitical instabilities to emerging digital threats that demand timely, accurate, and actionable risk assessments. Moreover, modern supply chains consist of hundreds or even thousands of suppliers, each generating a vast amount of news and textual data that must be analyzed for potential risk events. This immense volume and complexity of information render traditional risk identification techniques, which often rely on labor intensive expert analyses and conventional data-driven methods, increasingly inadequate.

Recent advances in Artificial Intelligence (AI) and, more specifically, the emergence of Large Language Models (LLMs) have ushered in transformative potential for SCRM. LLMs are adept at processing and contextualizing vast amounts of unstructured text, from vendor contracts and industry reports to real-time news feeds, enabling them to extract relevant risk signals that might otherwise go undetected. Models such as OpenAI's GPT series and Anthropic's Claude series not only generate human-like summaries of complex information but also provide dynamic, interactive risk assessments that support proactive decision making.

Building on this technological momentum, our prior work introduced the LARD-SC framework [2, 3], a novel, service-oriented solution tailored for supply chain risk identification and analysis. LARD-SC leverages advanced prompt engineering, text embeddings, and graph-based visualizations to streamline the entire risk management life cycle. It automates the collection of real-time supplier news, interprets the extracted textual data using LLMs, and classifies risks using a hierarchical taxonomy based on *Class*, *Family*, and *Type*. By integrating these service components with the intuitive Neo4j data visualization, LARD-SC offers a comprehensive and user-friendly approach that empowers risk managers (RMs) to navigate the intricate web of supply chain vulnerabilities.

This paper extends our previous efforts by conducting an additional benchmarking study on various state-of-the-art LLMs for risk identification within supply chain contexts. We systematically evaluate five widely adopted LLM variants: GPT-3.5 turbo, GPT-4o, GPT-4o mini, Claude 3.5 Sonnet, and Claude 3.5 Haiku, using a consistent prompt-based methodology. Through quantitative assessments based on expert reviews, we derive performance metrics such as the Risk Validation Rate (RVR), Potential Risk Rate (PRR), and False Identification Rate (FIR), culminating in a comprehensive Relative Performance Index (RPI) for each model. Our evaluation of 120 news samples related to Apple's Tier-1 suppliers reveals that while all models show promising capabilities, newer generation LLMs such as the GPT-4o variant exhibit superior performance in accurately identifying and assessing supply chain risks, thereby addressing the critical need to process large volumes of supplier-related news efficiently and reliably.

By integrating cutting-edge AI capabilities with robust risk management methodologies, this work aims to contribute to the evolution of SCRM into a more agile, service-oriented, data-driven discipline capable of addressing the dynamic challenges of today's global supply chains. The primary contributions of this paper are:

1. **Comprehensive LLM Benchmarking:** Leveraging a consistent and robust evaluation methodology, we benchmark multiple LLMs on their ability to extract, summarize, and assess supply chain risk from diverse textual sources. Our study uses a carefully structured prompt system and expert validations to ensure that the results are both reproducible and reflective of real-world decision making contexts.
2. **Empirical Insights for Risk Management:** By deriving key performance metrics (such as RVR, PRR, and FIR) through expert validation, the paper provides tangible evidence on how and why recent advanced LLMs outperform earlier models in identifying potential disruptions across diverse supply chain contexts.
3. **Enhanced Risk Identification via GPT-4o:** We improve the risk identification process by integrating high-performing LLMs into the LARD-SC framework. This integration demonstrates a significant enhancement in identifying and assessing potential supply chain risks compared to our initial GPT-3.5 turbo model, resulting in more accurate risk detection and timely actionable insights.

The remainder of this paper is organized as follows. Sect. 3 reviews the related literature and positions our work within the broader context of supply chain risk identification and AI applications. Sect. 4 details the LARD-SC framework, outlining its various service components and the technological underpinnings that support effective risk analysis. Sect. 5 presents the risk identification benchmark process, detailing the methodology, evaluation metrics, and consolidated findings drawn from multiple expert reviews and Sect. 6 concludes with a summary of key findings and directions for future research.

3 Background and related work

SCRM has become an essential discipline in supply chain operations due to the widespread uncertainties and disruptions that may affect the smooth flow of products and services [4, 5]. At its core, risk identification involves the systematic detection and assessment of potential threats, forming the basis for subsequent mitigation strategies [6]. Recent advancements in AI, and more specifically LLM, have opened new frontiers in risk identification. LLMs such as GPT [7] and Claude [8] are capable of parsing large volumes of unstructured text, detecting early risk signals, synthesizing domain-specific insights, and providing actionable recommendations [9]. This section reviews the evolution of risk management techniques from qualitative and quantitative methods to AI-based approaches that leverage advanced language understanding.

3.1 Supply chain risk management: An overview

Supply chain risks arise from disruptions at any stage of the value chain, from raw material procurement to final distribution, and encompass factors such as demand fluctuations, supplier failures, logistics bottlenecks, and geopolitical instability [10, 11]. Traditionally, risk identification has employed qualitative methods (e.g., brainstorming, Delphi techniques) and quantitative approaches (e.g., statistical forecasting and simulation). However, these methods can be resource intensive and heavily reliant on expert judgment [5]. Historically, risk identification in supply chains has relied on a combination of qualitative methods (e.g., brainstorming, Delphi,

scenario planning) and quantitative methods (e.g., statistical forecasting, simulation). However, these methods can be resource intensive and heavily reliant on expert knowledge. As global supply chains grow increasingly complex and data-rich, researchers and practitioners have turned to AI-based techniques to automate or augment parts of the risk identification process [12, 13].

3.2 AI-Based approaches in supply chain risk management

Prior to the emergence of LLMs, machine learning (ML) had already shown promise in detecting anomalies, classifying risk events, and forecasting disruptions in large data sets [14, 15]. These ML methods often rely on structured numerical data, such as historical lead times, shipping delays, or performance metrics, and can generate probabilistic risk scores. Techniques such as random forests for classification and outlier detection [14], and neural networks for predictive analytics and demand forecasting [16] demonstrate how data-driven approaches can improve supply chain visibility and support proactive decision making. Yet, while ML approaches offer notable accuracy gains, they typically require extensive feature engineering and are less adept at contextual reasoning, especially for unstructured text data [15]. This limitation has led to a growing interest in more advanced AI frameworks that can handle the linguistic nuances and semantic complexity of textual information.

3.3 LLMs in supply chain risk management

LLMs, such as GPT-4o [7], and Claude 3.5 [8], represent a new frontier in AI-driven text understanding. Building on transformer-based architectures [17], LLMs are trained on extensive corpora of internet text, enabling them to perform contextual analysis of risk-related documents (e.g., vendor contracts, industry reports, geopolitical news) to identify potential disruptions [9]. Leveraging their state-of-the-art natural language processing and ML capabilities, LLMs offer promising solutions for enhancing situational awareness, and optimizing decision-making processes [18]. Moreover, combining LLMs and domain expertise for predictive analytics enables robust supply chain disruption predictions, effectively addressing limitations of traditional methods such as interpretability and accuracy [19]. Generative AI, including LLMs, facilitates risk detection, analysis, and communication within supply chains [20].

Recent investigations emphasize the substantial promise of generative AI for supply chain and operations management. Gaurav [20] highlights the importance of using LLMs to identify, categorize, and prioritize events in supply chains. The study also underscores that integrating LLMs into SCRM systems can significantly enhance their effectiveness by addressing both known and unknown risks, enabling businesses to anticipate, absorb, adapt, and recover from disruptions. Jackson et al. [21] propose a capability-based framework that details how organizations can effectively implement LLMs and other generative technologies in various stages of the supply chain. Similarly, Zheng et al. (2024, preprint) highlight the practical potentials of LLMs for risk identification through a dedicated case study, demonstrating how these models can streamline data collection, synthesize unstructured information, and provide more proactive mitigation strategies. Cheng et al. [19] leverage fine-tuned LLMs and Graph Convolutional Networks (GCNs) to extract, analyze, and predict disruptions with a human-in-the-loop approach, showcasing its effectiveness through evaluations on real-world data. Kuhl et al. [22] demonstrates the importance of optimizing LLM parameters, such as temperature and top P, to enhance supply chain risk detection.

Building upon these insights, our previous works focus specifically on leveraging LLMs for SCRM. In the first study [2], we introduced the LARD-SC framework, an integrated framework for optimizing SCRM processes with LLMs, illustrating how LLM-driven analytics can expedite risk detection and enhance mitigation pathways. Subsequently, we further developed software prototypes and interactive visualization tools to showcase how LLM-augmented workflows can support decision making across complex supply networks [3]. These foundational contributions underscore the critical role of advanced text analytics in uncovering hidden vulnerabilities within supply chains. Despite their promise, LLMs also introduce new challenges. For instance, they can sometimes produce incorrect or fabricated information [9]. In a high-stakes domain like SCRM, reliance on unverified outputs can be risky. Moreover, integrating LLMs into decision support systems raises questions about data privacy, fairness, and responsibility for actions taken based on AI-generated insights [23]. Finally, running large-scale LLMs in real-time can be resource-intensive, and many organizations may lack the infrastructure or budgets to implement such systems [24]. Nonetheless, the expanding body of literature on generative AI and LLM implementations in supply chain management underscores the technology's considerable potential to transform traditional risk identification processes. This paper thus builds upon these prior endeavors to offer a more comprehensive examination of how LLMs can be practically deployed to enhance supply chain resilience.

4 LARD-SC framework overview

As introduced in the previous study [2], the LARD-SC framework represents an innovative approach to managing risks within supply chain networks (SCNs), addressing the multifaceted challenges faced by organizations in monitor-

ing and mitigating supplier risks. This framework integrates advanced technologies and methodologies to provide a comprehensive, efficient, and user-friendly service for proactive risk management. LARD-SC is an overarching framework that develops and integrates three sub-frameworks, namely DCV-RA (Data Collection and Visualization for Risk Analysis), LLM-RI (Large Language Model-based approach for Risk Identification) and LLM-RC (Large Language Model-based approach for Risk Classification) for automated risk identification, risk classification, and interactive visualization of the detected risks. Fig. 1 provides an overview of how the three sub-frameworks collectively address the spectrum of LARD-SC's requirements for proactive supply chain risk identification and management.

4.1 DCV-RA: Data collection and visualization for risk analysis

DCV-RA is the first sub-framework of the LARD-SC framework that presents the risk information intuitively and interactively. This interactive functionality empowers managers to navigate vast data sets effectively, fostering better alignment between risk assessments and organizational strategies. The LARD-SC framework uses the DCV-RA sub-framework in two phases, namely the pre-analysis phase and the post-analysis phase.

4.1.1 Pre-analysis phase

In its pre-analysis phase, DCV-RA leverages the Neo4j graph database and its associated Neo4j Browser to construct and visualise a preliminary representation of the focal company's SCN. This phase is crucial for establishing a foundational understanding of supplier relationships and their geographical distribution. The selection of a graph database is predicated on its inherent capacity for flexible data modeling, a characteristic particularly advantageous for representing the intricate and often dynamic nature of modern SCNs. Graph databases, unlike traditional relational models, naturally accommodate the complex interconnections and evolving relationships inherent in supply chain ecosystems. Furthermore, the DCV-RA sub-framework incorporates the GoogleNews Python library, alongside other specialized data ingestion tools, to ensure the continuous and automated population of the system with up-to-date, supplier-relevant news articles. This mechanism is configured to systematically query and retrieve articles that explicitly reference each supplier name within the defined supplier database. This automated retrieval process is crucial for ensuring that the system maintains a temporally relevant and up-to-date repository of relevant articles. By continuously and autonomously scanning for newly published content, the subsequent LLM-RI sub-framework proactively captures evolving information relevant to supplier risks. Upon the successful retrieval of articles, the system initiates a sequence of essential preprocessing stages to ensure data quality. These stages include: the execution of deduplication algorithms to eliminate redundant entries arising from multiple news sources covering the same event; subsequently, employing libraries like Beautiful Soup, ensuring the HyperText markup language (HTML) structure is parsed and cleaned, to extract the core article text, effectively removing extraneous HTML tags, scripts, and navigational elements. This preprocessing stage ensures that only the essential textual content is passed to the subsequent LLM-based analysis, enhancing both the accuracy and efficiency of risk identification; and finally, the structured organization and persistent storage of the resultant, refined data is organized within a dedicated database table, in a local relational database environment. This data acquisition and integration process in the pre-analysis phase is essential for providing a timely and contextually rich dataset upon which subsequent advanced risk analytics are predicated.

4.1.2 Post-analysis phase

Subsequent to the pre-analysis phase, the DCV-RA sub-framework transitions into its post-analysis phase, wherein it systematically captures granular risk details and standardized classifications generated by the LLM-RI and LLM-RC sub-frameworks. These analytically derived insights are then seamlessly appended to the dynamic visualization network within Neo4j, enriching the graphical representation of the supply chain risk landscape. This integration empowers RMs to effectively leverage the resultant risk events and their associated CTBR risk labels as actionable intelligence. Through the interactive Neo4j Browser, RMs can filter, explore, and meticulously scrutinize risk events, enabling nuanced analysis by geographical region, specific supplier, or predefined risk category. These analytical operations are facilitated by real-time graph representations, which dynamically reflect the most current risk data and interrelationships. Furthermore, in this phase, the DCV-RA sub-framework integrates complementary metadata originating from the news sources, thereby providing a comprehensive audit trail of evidence for each identified risk. This feature is paramount for enhancing the transparency and verifiability of risk assessments. For enhanced decision-making support, the DCV-RA sub-framework is engineered to present RMs with readily accessible key metadata elements. These include direct links to the primary source article from which a risk was derived, the LLM-generated rationale underpinning assessments of high risk likelihood or impact, and the standardized CTBR risk category assigned to each event. This consolidated presentation of evidence and analytical justification significantly augments the capacity of RMs to make informed, data-driven decisions regarding risk mitigation and strategic response.

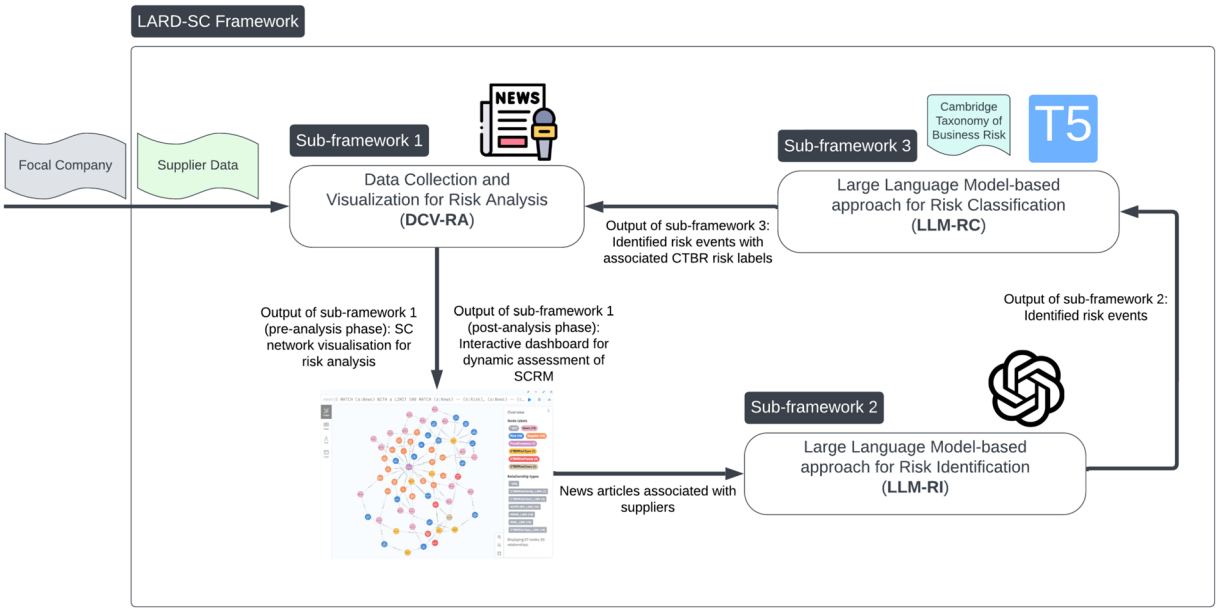

Fig. 1 Conceptual Model of the LARD-SC Framework. Integrates DCV-RA, LLM-RI, and LLM-RC to enable automated risk identification, classification, and visualization for proactive SCRM

For instance, to facilitate a targeted analysis of the risks associated with a particular risk type, such as *Product Defect/Failure*, RMs can readily utilize the predefined Cypher query template to retrieve only those nodes and relationships within the graph database that are directly associated with the *Product Defect/Failure* risk type, effectively filtering the visualization to focus on this specific risk category. This targeted query simplifies the resultant graph visualization, making it considerably more focused and directly relevant to the user's specific analytical needs. The filtered visualization, generated by executing this customized query, is presented in Fig. 2, showcasing the enhanced clarity and interpretability achieved through targeted data filtering and query customization.

Furthermore, RMs can seamlessly navigate through the visualized risk network, dynamically exploring the properties of each node and examining its direct and indirect connections to other nodes, enabling in-depth data drill-down and contextual risk analysis. For instance, RMs can select and inspect a `Risk` node within the Neo4j Browser to access a comprehensive set of properties associated with that specific risk event, such as risk description, risk likelihood/impact and its rationale. Additionally, RMs can further review the corresponding `News` node to examine the source information and generated summary associated with the news article that triggered the risk event.

Figures 3 and 4 illustrate the properties panel of a selected `Risk` node and a `News` node within the Neo4j Browser, respectively, showcasing the readily accessible detailed risk or source information.

By following the labeled edges that dynamically connect nodes within the graph visualization, RMs can effectively trace the complex relationships and interdependencies between suppliers, identified risks, and the originating news articles. This interactive graph exploration capability empowers RMs to gain a comprehensive and contextually rich understanding of how specific events, as reported in news articles, may potentially impact individual suppliers and propagate throughout the broader SCN, facilitating informed risk assessment and targeted mitigation planning.

4.2 LLM-RI: large language model-based approach for risk identification

LLM-RI is the second sub-framework of the LARD-SC framework. This sub-framework takes the retrieved news articles associated with each supplier and uses an LLM module to examine these articles. This will isolate potential disruptions, offering a likelihood estimate and an impact assessment for each detected risk event. As shown in Fig. 1, the output of this sub-framework assists the RMs of a focal company in visualizing its geographically spread SCN with the determined risks impacting it and its suppliers with their impact.

Following the ingestion of these news articles, the LLM-RI sub-framework uses a specialized large language model prompt and function-calling strategy to derive structured outputs from textual inputs. For each article, it captures two essential dimensions: likelihood, which is an estimate of the probability or frequency of the event's occurrence and

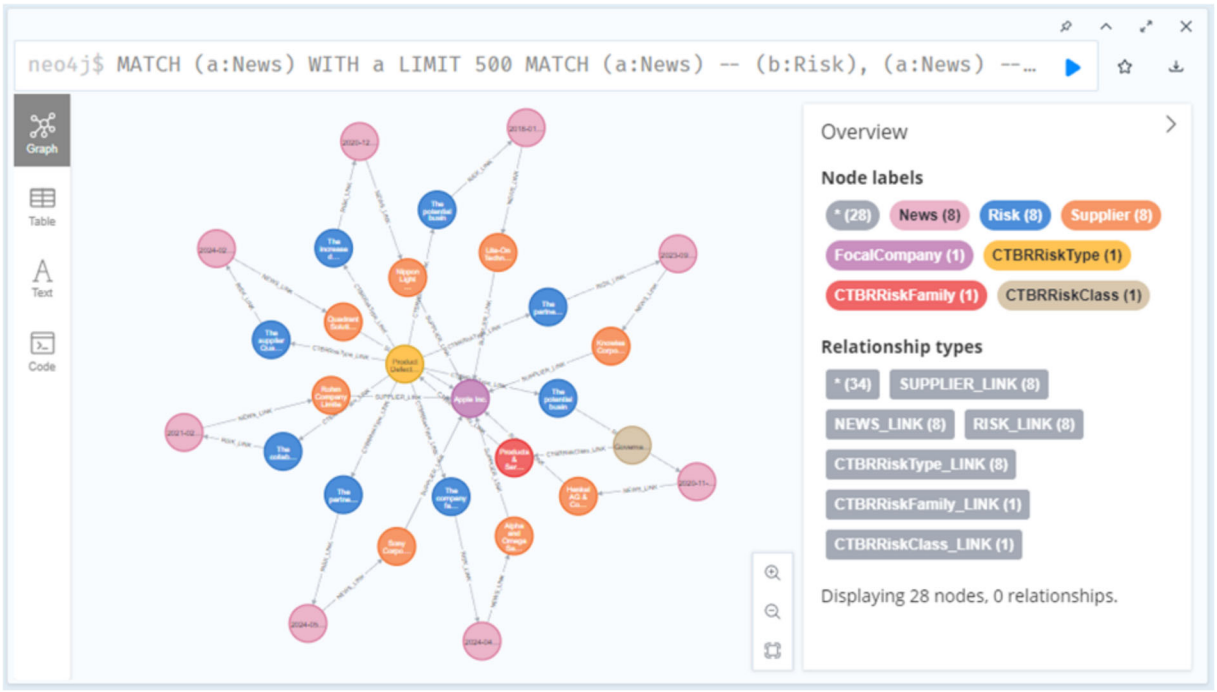

Fig. 2 Filtered Visualization by Risk Type. Users can drill down into specific risk types (e.g., *Product Defect/Failure*) to analyze how they propagate through the network

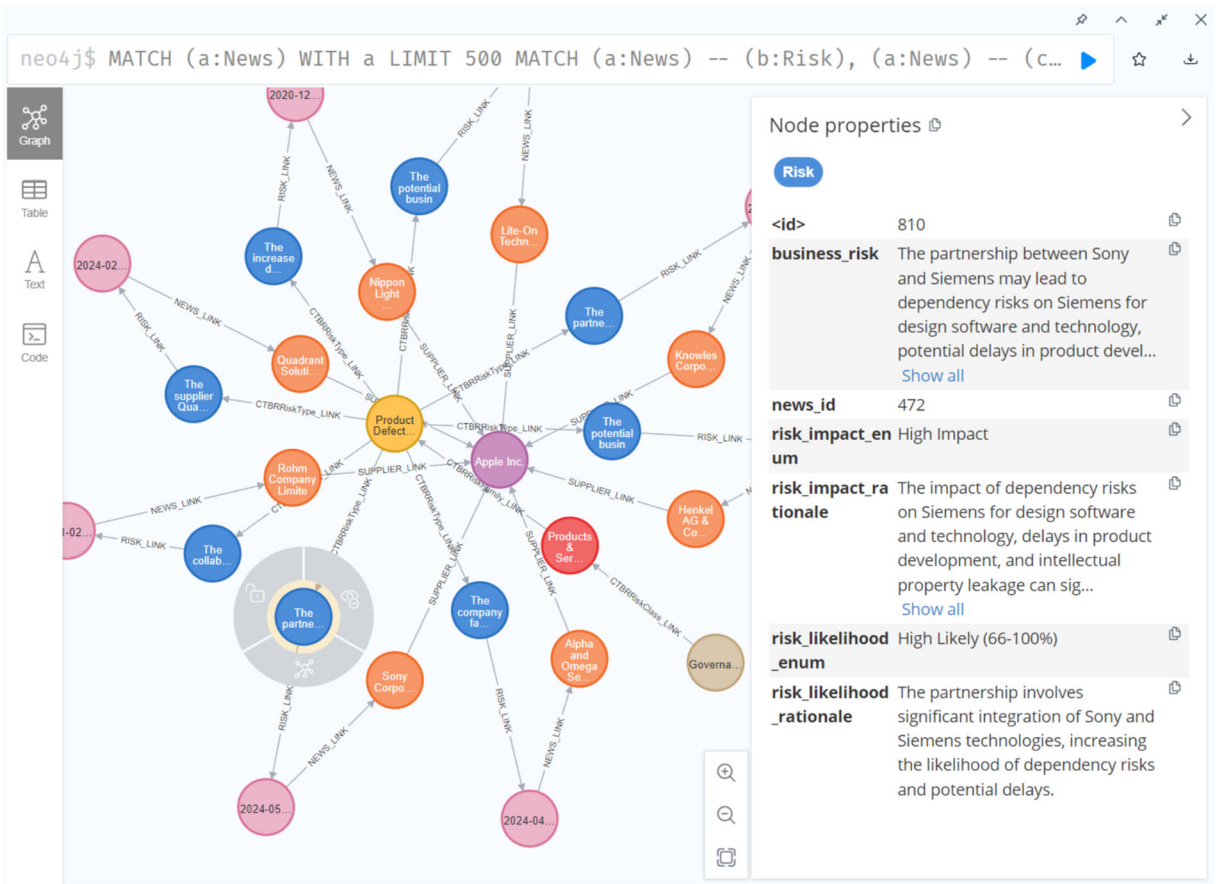

Fig. 3 Detailed View of a Risk Node's Properties. This interface provides quick access to the news summary and the model's likelihood/impact assessments

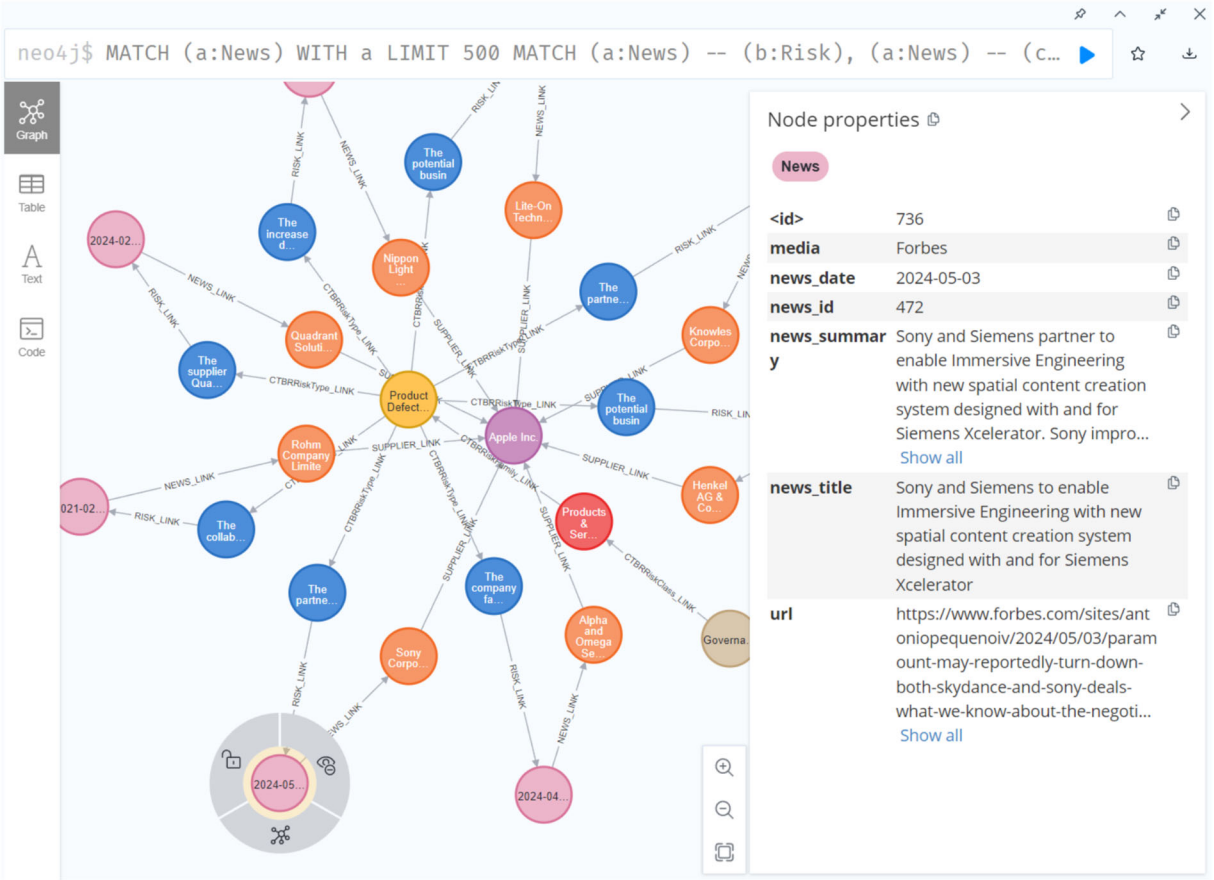

Fig. 4 Detailed View of a News Node's Properties. This interface provides quick access to the original news article with additional metadata

impact, which is a measure of the event's potential adverse consequences on supply chain performance, operations, or stakeholder welfare. By integrating these two dimensions, the LLM-RI sub-framework automatically flags high-risk events.

By coupling advanced LLM capabilities with outputs from the automated news collection module, LLM-RI provides a proactive and scalable solution for capturing early warning signals of potential supply chain vulnerabilities. This proactive stance is crucial for enabling timely intervention and mitigating the adverse impacts of disruptions. LLM-RI is built upon the seamless integration of OpenAI's state-of-the-art GPT-4o LLM via the OpenAI API. This integration of a cutting-edge LLM is central to the advanced analytical capabilities, enabling the system to harness the most current and sophisticated NLP techniques for systematic supply chain risk analysis. By leveraging the GPT-4o model, the system effectively capitalizes on its exceptional ability to accurately comprehend and fluently generate human-like text, rendering it particularly adept at navigating the inherent linguistic complexities and nuances that are characteristic of real-world risk evaluation scenarios. The GPT-4o model's interpretative functions are instrumental in enabling the LARD-SC framework to not only reliably detect explicit risk signals embedded within unstructured news articles but also to intelligently capture subtle, context-specific cues that are often indicative of emerging or latent supply chain threats. This advanced LLM-powered risk identification module thus facilitates a proactive, highly nuanced, and exceptionally efficient approach to SCRM, ensuring that potential risk events are identified with both high precision and operational efficiency.

To effectively guide the LLM in accurately extracting relevant risk information from unstructured news articles and consistently generating structured, actionable outputs, the LLM-RI employs prompt engineering techniques. Prompt engineering, in this context, is not merely about providing instructions to the LLM; it is a critical methodological component for effectively directing the model's analytical attention to specific SCRM tasks and ensuring that the LLM-generated responses are consistently accurate, contextually relevant, and practically actionable for risk management decision-making. The prompt design is iterative and empirically refined to optimize the LLM's performance in extracting relevant risk information. A typical prompt sequence employed in LLM-RI comprises three key components:

1. **System Context Setting:** The prompt begins by establishing the system context, explicitly defining the LLM's role and perspective. This is achieved through a system-level instruction that sets the LLM's persona as a professional risk assessor operating within the context of the *{Focal_Company}*. This contextualization is crucial for aligning the LLM's reasoning with the specific objectives of SCRM. The system context prompt is defined as: *You are a professional risk assessor for {Focal_Company}.*
2. **Step-by-Step Instructions:** To guide the LLM's analytical process, a series of explicit, step-by-step instructions are provided within the prompt. These instructions decompose the complex task of risk identification into a sequence of manageable sub-tasks, ensuring a structured and systematic approach. The instructions are designed to elicit specific information from the LLM, including text summarization, relevance assessment, risk identification, and risk assessment. The detailed steps are as follows:
    - *Step 1: Summarize news content.* This step instructs the LLM to generate a concise summary of the provided news article, capturing the key information and events described within the text.
    - *Step 2: Check whether the news is related to {Supplier_Name}.* This step directs the LLM to assess the relevance of the news article to the specified supplier, ensuring that the subsequent risk analysis is focused on supplier-specific disruptions. A textual rationale to justify its assessment of the news's relevance to the supplier is also supplemented.
    - *Step 3: If related, identify potential supply chain risks, specifying relevant details.* Conditional upon the news being deemed relevant to the supplier, this step instructs the LLM to identify and articulate potential supply chain risks that are indicated by the news content. The LLM is further instructed to provide specific details pertaining to the identified risks, such as the nature of the risk event and the affected aspects of the supply chain.
    - *Step 4: Evaluate risk likelihood and impact.* Finally, this step prompts the LLM to evaluate the identified risks in terms of their likelihood of occurrence and potential impact on the supply chain, and provide the rationale or justifications. This risk assessment component is crucial for prioritizing risks and informing subsequent risk mitigation strategies.
3. **User-Supplied Contextual Input:** The final component of the prompt sequence comprises the user-supplied contextual input, which provides the LLM with the specific data it needs to perform the risk identification task. This input includes the preprocessed article text, the supplier name associated with the article, and any other relevant metadata that may enhance the LLM's understanding of the context. This dynamic input ensures that the prompt is tailored to each individual article-supplier pair, enabling context-aware risk analysis.

The use of structured prompts and function calling techniques ensures that the language model returns data in a format suitable for analysis and visualization, allowing for comprehensive and consistent risk assessments throughout the supply chain.

To illustrate the granular nature of risk insights generated by LLM-RI, Fig. 5 presents a representative example of a risk event identified from news collected for Apple supplier 3M. The news article, titled *3M Announces Departure of Chief Financial Officer* [1], was processed by LLM-RI, resulting in the identification of a potential supply chain risk event characterized by the following description:

> *Potential disruption in financial leadership and decision making at 3M. Possible delays or shifts in financial strategies, budgeting, or investments that could affect supply chain stability. Market confidence in 3M might be affected, influencing stock prices and financial health, which could impact its operational capacity and reliability as a supplier.*

Furthermore, LLM-RI provided a nuanced risk assessment, evaluating both the likelihood and potential impact of this event:

- **Risk Likelihood:** *Moderate Likelihood.* Rationale: *Leadership transitions in large companies like 3M are common, but they can introduce short-term uncertainties in financial planning and operational decisions.*
- **Risk Impact:** *Moderate Impact.* Rationale: *While 3M is a well-established company with robust systems, the CFO departure may cause temporary instability or adjustments that could affect pricing, supply timelines, or overall reliability in the short-to-medium term.*

This example demonstrates LLM-RI's capability to not only identify potential risk events from seemingly high-level corporate news but also to provide a structured and reasoned assessment of their potential implications for supply chain operations. The system effectively translates a news item about executive leadership change into a tangible supply chain risk, complete with likelihood and impact evaluations.

[1] https://www.prnewswire.com/news-releases/3m-announces-departure-of-chief-financial-officer-302192682.html

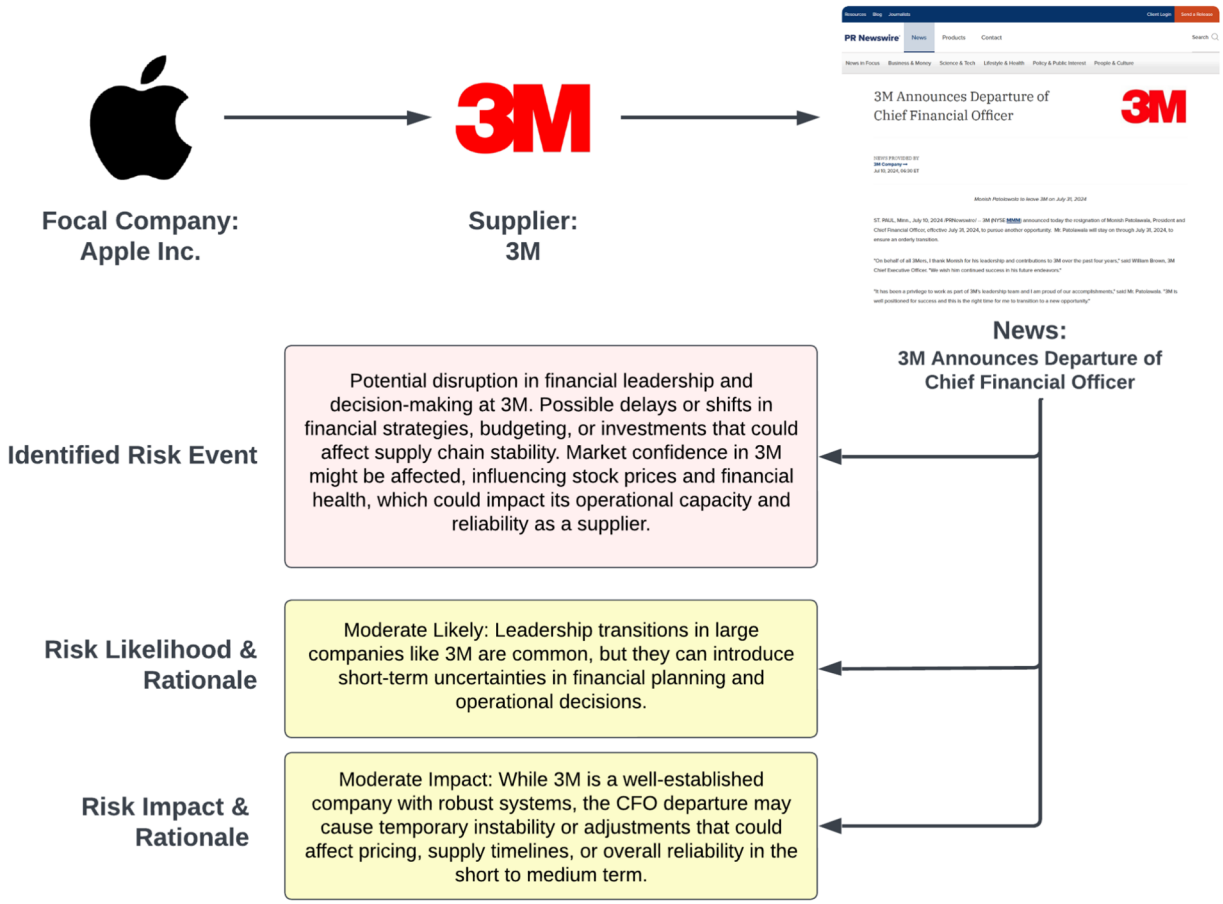

Fig. 5 Example of LLM-RI Output: Risk Identification from News of 3M CFO Departure. The system identifies the potential risk and provides likelihood and impact assessments

4.3 LLM-RC: Large language model-based approach for risk classification

After LLM-RI identifies potential risks, LLM-RC, the final sub-framework of the LARD-SC framework, employs advanced embedding techniques and the CTBR to classify these disruptions into coherent risk categories that may influence the focal company's operations. Organizing risk events by category allows RMs to focus on specific threat domains (e.g., financial instability, geopolitical uncertainties, environmental contingencies).

The LLM-RC sub-framework assists in the classification of the identified risk events according to the focal company's interests. Free-text descriptions of risk can hinder effective decision making if they are not systematically organized. To have a structured representation of the risks identified by LLM-RI, the LLM-RC sub-framework uses the CTBR that offers a well-established framework for categorizing a wide spectrum of financial, geopolitical, technological, environmental, social, and governance risks [1]. By using such a taxonomy, the LLM-RC sub-framework employs LLMs with embedding-based techniques (such as sentence-t5-base) to map textual descriptions of identified risks to the nearest CTBR category. In this manner, the classification process becomes more context-aware and aligns unstructured risk narratives with a consistent, standardized taxonomy. This approach significantly enhances the interpretability and comparability of risk information across different sections of the supply chain. This output is given to the DCV-RA sub-framework, an interactive visualization layer presenting consolidated insights, comprising risk assessments and risk labels to RMs. The LLM-RC sub-framework thus assists the RM to generate actionable insights for prompt and well-informed decision making in SCRM.

Risk events identified and extracted by systems like LLM-RI are frequently articulated as free-text descriptions. For instance, an event might be described as *"supplier experiencing shortages of essential raw materials due to unforeseen natural disaster."* While such textual descriptions are rich in detail and contextually informative, their unstructured nature poses significant challenges for effective risk management. Specifically, free-text risk descriptions are not directly amenable to quantitative comparison, aggregation across different events, or systematic trend analysis. Furthermore, the

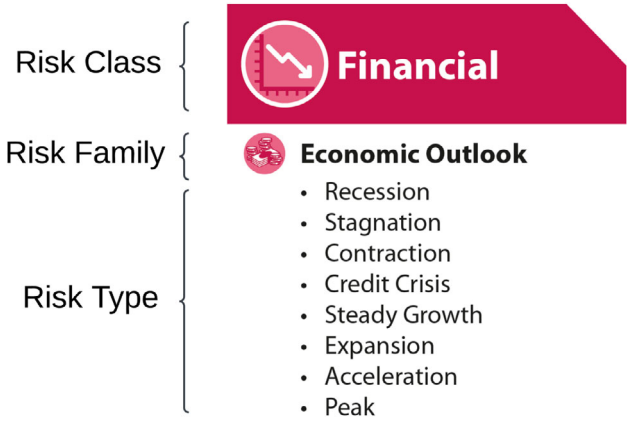

Fig. 6 Hierarchical Structure of the CTBR. Organizes business risks into multi-level taxonomy: Class, Family, and Type

subjective interpretation of such events can vary considerably among RMs unless a standardized and universally applied reference framework, such as a risk taxonomy, is consistently employed. This lack of structure impedes the development of a cohesive and organization-wide understanding of risk exposure.

To address the limitations of unstructured risk data, the integration of a robust risk taxonomy is paramount. The CTBR emerges as a particularly suitable framework. CTBR is a widely recognized hierarchical taxonomy that systematically structures business risks into six overarching top-level `Class`. Each class is further decomposed into multiple `Family`, and subsequently, into more granular `Type`, providing a multi-layered classification structure, as shown in Fig. 6.

Subsequent to the detection and extraction of unstructured risk events by LLM-RI, the LLM-RC sub-framework undertakes the critical task of systematically classifying these initially unstructured events into a structured taxonomy. The core operational mechanism of LLM-RC, centers on the application of text embedding and semantic similarity techniques. In this workflow, the *sentence-t5-base* LLM is employed to generate dense vector embeddings for both the identified unstructured risk event description and the set of CTBR labels (e.g., *Financial + Economic Outlook + Recession*, exemplified here using the *Class + Family + Type (CFT)* textual representation approach, a method validated as superior in accuracy compared to alternative approaches in a previous study [2]. Following the embedding generation phase, the system proceeds to calculate the semantic similarity score between the generated risk event embedding and each individual CTBR label embedding, employing the cosine similarity metric as the quantitative measure of semantic relatedness. Upon completion of the pairwise semantic similarity computations across all CTBR labels, the resulting scores are systematically ranked in descending order. The CTBR label exhibiting the highest computed semantic similarity score is then assigned to the input risk event, thereby completing the classification process.

For instance, in a representative classification scenario, the CTBR label *Social - Human Capital - Loss of Key Personnel* will be assigned to this specific risk event *"The management changes and restructuring efforts may impact employee morale and could lead to the loss of key talent if not managed carefully."* with a high semantic similarity score of 0.811258, indicating a strong semantic alignment. Ultimately, LLM-RC effectively maps each unstructured risk event, extracted from textual sources by LLM-RI, to the most semantically appropriate *class*, *family*, and *type* categories as defined within the hierarchical structure of the CTBR. This comprehensive mapping process fundamentally transforms the initially unstructured risk descriptions into structured, taxonomically organized risk intelligence, thereby enabling downstream analytical and decision-making functionalities.

4.4 LARD-SC software prototype

The LARD-SC software prototype represents the tangible implementation of the overarching LARD-SC framework, effectively operationalizing the conceptual model for practical application in SCRM. This prototype serves as an end-to-end software system that integrates the functionalities of the DCV-RA, LLM-RI, and LLM-RC sub-frameworks, demonstrating their synergistic capabilities in a cohesive and functional tool. The primary objective of the LARD-SC software prototype is to automate and streamline the entire SCRM process, from initial data ingestion and real-time news collection to advanced risk identification, structured classification, and interactive visualization. By providing a working instantiation of the LARD-SC framework, the prototype validates its feasibility and showcases its potential to enhance proactive and data-driven SCRM practices.

The LARD-SC software prototype is designed with a modular architecture, comprising five core modules that mirror the workflow of the LARD-SC framework. These modules, namely *Initial Configuration*, *Automated News Collection*, *Risk Identification and Assessment*, *Risk Classification*, and *Data Visualization*, operate interdependently to deliver an end-to-end SCRM solution. The prototype leverages a combination of technologies, including SQLite and Neo4j databases, OpenAI's GPT-4o LLM, and the `sentence-t5-base` embedding model, to achieve its functionalities. The *Initial Configuration* module sets up the system environment and ingests supplier data, establishing the foundation for subsequent processes. The *Automated News Collection* module continuously retrieves supplier-specific news articles from Google News, providing up-to-date data input for risk analysis. The *Risk Identification and Assessment* module employs GPT-4o to analyze news

articles, identify potential risk events, and assess their likelihood and impact. The *Risk Classification* module utilizes the `sentence-t5-base` model and CTBR taxonomy to categorize identified risks into a structured hierarchy. Finally, the *Data Visualization* module leverages Neo4j to present an interactive graph-based dashboard, enabling RMs to explore and analyze the integrated risk data effectively.

4.5 Advantages and applications of the LARD-SC Framework

The LARD-SC framework offers several advantages for SCRM, including automation capabilities that reduce the time and effort required to monitor and analyze risks, the integration of advanced language models that ensure high precision in risk identification and classification, and graph-based visualization that improves the accessibility and interpretability of complex data, allowing organizations to make data-driven decisions. This framework is applicable across various industries, from manufacturing to retail, where supply chain risks pose significant challenges. Its scalability and modular design make it adaptable to organizations of different sizes and complexities. By providing a comprehensive service component for risk management, the LARD-SC framework empowers organizations to safeguard their supply chains and maintain operational resilience.

In summary, the LARD-SC framework represents a significant advancement in SCRM, integrating cutting-edge technologies and methodologies to deliver a powerful, intuitive, and efficient service for identifying, classifying, and mitigating supply chain risks.

5 Risk identification benchmark

To comprehensively assess the efficacy of LLM-RI in real-world supply chain risk identification scenarios, a comprehensive benchmarking study was conducted. This evaluation encompasses both quantitative metrics and qualitative expert review, providing a multifaceted assessment of LLM-RI's performance. The evaluation benchmarks multiple state-of-the-art LLMs.

This section presents an extended evaluation of the LARD-SC framework to manage supply chain risks through the lens of multiple LLMs. Building on our earlier studies, we test five widely used LLM variants from both OpenAI and Anthropic:

- `GPT-3.5 turbo`: OpenAI's widely adopted and cost-effective model, utilized in our previous study, serving as a baseline for comparison [25].
- `GPT-4o`: OpenAI's new flagship model, representing a significant advancement in capabilities, including enhanced reasoning, contextual understanding, and task-solving abilities [26]. Available in both original (`gpt4o`) and smaller (`gpt4o_mini`) versions.
- `GPT-4o mini`: A smaller, more computationally efficient variant of GPT-4o, designed for applications requiring lower latency and resource consumption [27].
- `Claude 3.5 Sonnet`: Anthropic's highly capable model from the Claude 3.5 family, known for its strong language understanding, generation capabilities, and focus on safety and transparency [28].
- `Claude 3.5 Haiku`: A smaller and faster variant of Claude 3.5 Sonnet, optimized for speed and efficiency while maintaining strong performance [29].

GPT-4o and Claude 3.5 represent the newer generation of LLMs, incorporating architectural innovations and training methodologies that significantly enhance their performance compared to previous models like GPT-3.5 turbo. Including these models in the benchmark enables a robust evaluation of state-of-the-art LLM performance in the specific domain of supply chain risk identification.

The primary objective of this multi-LLM benchmark is to empirically compare these models across real-world supply chain risk scenarios, culminating in the identification of the top-performing LLM based on a rigorous evaluation methodology. This comparative analysis provides valuable insights for model selection and informs the practical deployment of LLM-RI in diverse operational contexts.

5.1 Inputs

To ensure a fair and rigorous comparison across different LLMs, the experimental setup was carefully designed to maintain consistency in input data and prompting strategies. A sample dataset of 120 news articles, specifically relevant to Apple's Tier-1 suppliers, was selected for this benchmark. This dataset was sampled from a dataset comprising 676 news articles collected through DCV-RA sub-framework, pertaining to 188 Apple Tier-1 suppliers. Apple Inc. was chosen as the focal company due to the extensive public information available regarding its supply chain and the criticality of SCRM for its global operations. The news articles were autonomously fetched over a defined time window spanning from 2020 to 2024. This dataset was curated to ensure coverage of a diverse range of supply chain risk topics, including logistics delays, financial issues, environmental compliance concerns, labor disputes, and geopolitical risks. This diversity in risk topics ensures that the benchmark evaluation is representative of the multifaceted nature of real-world supply chain risks.

All 120 news articles are ingested into the candidate LLMs following the LLM-RI workflow to identify potential risk signals from the news among Apple's Tier-1 suppliers. Crucially, the prompts and instructions provided to each LLM were kept strictly consistent across all models. This standardization of prompting is essential for isolating the performance differences attributable to the LLM architectures themselves, rather than variations in prompting strategies. To ensure reproducibility, LLM parameters such as temperature setting is set to zero, effectively eliminating stochastic variations in token generation to ensure a consistent output. Both OpenAI and Anthropic APIs support function-calling features, which were leveraged to ensure structured output retrieval from all LLMs. The standardized prompt employed in this study is detailed below:

1. **Role: System** - You are a professional risk assessor working at the focal company: {focal_company}.
2. **Role: System** - Use the following step-by-step instructions to respond to user inputs. If the answer cannot be found in the articles, output 'N/A'. If the answer cannot be found in the articles, output 'N/A'. The user will provide you with text in triple quotes.
3. **Role: System** - Step 1 - Summarize the news content.
4. **Role: System** - Step 2 - Determine whether the provided news is related to the supplier: {supplier_name} with the rationale.
5. **Role: System** - Step 3 - If the result in the Step 2 is 'Related', then proceed; otherwise output 'N/A' for all following steps.
6. **Role: System** - Step 4 - Identify potential business risks related to the supplier:
7. **Role: System** - Step 5 - Evaluate the business risk result you identified in Step 4 and provide the risk assessment result.
8. **Role: User** - Supplier: {supplier_name} News content: {news_content}

By adhering to these standardised instructions, each LLM was tasked with providing a consistent and structured breakdown of news relevance and potential risk statements, facilitating direct performance comparison across models.

5.2 Evaluations

To evaluate the reliability and validity of LLM-RI's outputs, a multi-expert validation process was implemented, drawing upon the methodology of multi-expert voting to mitigate individual biases and enhance the robustness of the assessment. To ensure the robustness and reliability of the multi-LLM benchmark evaluation, a panel of four domain experts participated in the validation process. The expert panel comprised individuals with distinct but complementary areas of expertise: two experienced SCRM specialists and two artificial intelligence/natural language processing (AI/NLP) specialists. This balanced composition ensured that the evaluation considered both the practical relevance of identified risks and the technical capabilities of the LLMs.

For the purposes of evaluation, these four experts were randomly paired into two independent groups, designated as *Expert Pair A* and *Expert Pair B*. Each expert pair independently assessed the risk events generated by each of the five LLM variants, adhering to the same rigorous validation criteria employed in the single-LLM evaluation:

- **Source Validity:** This criterion assessed whether the identified risk event was directly and unambiguously derived from the provided news content. Experts were asked to determine if the LLM's risk identification was grounded in the information presented in the article.
- **Logical Reasonability:** This criterion evaluated the logical coherence and plausibility of the identified risk event as an interpretation of the news content. Experts were asked to judge whether the LLM's interpretation of the news as a risk event was logically sound and reasonable within the context of supply chain operations.
- **Impact Assessment Relevance:** This criterion focused on the relevance of the identified risk event in terms of its potential negative impact on the supplier in question. Experts were asked to assess whether the identified event demonstrably represented a clear and credible negative impact on the supplier's operations or financial stability.

For each risk event and each LLM, each expert within a pair cast binary (true/false) votes on each of the three validation criteria. A combined validity function, $V(r_i)$, was then assigned to each risk event r_i based on the aggregated votes from both expert pairs, as defined in Equation 1:

$$V(r_i) = \begin{cases} 2, & \text{if both expert pairs vote true } (\textit{Valid}), \\ 1, & \text{if one expert pair votes true } (\textit{Potentially Valid}), \\ 0, & \text{if both expert pairs vote false } (\textit{Invalid}). \end{cases} \tag{1}$$

This validity function, consistent with the single-LLM evaluation, captures the consensus level among experts regarding the validity of each identified risk event, enabling the derivation of key performance metrics for each LLM. Specifically, for each LLM and across the entire dataset of evaluated risk events, the following metrics were computed:

- **Risk Validation Rate (RVR):** The Risk Validation Rate (RVR) quantifies the proportion of identified risk events that are fully validated by both experts. It is calculated

as:

$$\text{RVR} = \frac{|\{r \in R \mid V(r) = 2\}|}{|R|} \tag{2}$$

where R represents the set of all risk events identified by LLM-RI, and $V(r)$ is the validity function that assigns a score of 2 for fully validated events. RVR provides a measure of the system's precision in identifying genuinely valid risk events with high confidence.

- **Potential Risk Rate (PRR):** The Potential Risk Rate (PRR) measures the proportion of identified risk events that receive partial validation, i.e., validation from only one expert. It is calculated as:

$$\text{PRR} = \frac{|\{r \in R \mid V(r) = 1\}|}{|R|} \tag{3}$$

PRR quantifies the rate of events that are considered potentially valid, indicating cases where there is some level of expert agreement but also some ambiguity or disagreement. These events may warrant further investigation or closer scrutiny.

- **False Identification Rate (FIR):** The False Identification Rate (FIR) represents the proportion of identified risk events that are deemed invalid by both experts. It is calculated as:

$$\text{FIR} = \frac{|\{r \in R \mid V(r) = 0\}|}{|R|} \tag{4}$$

FIR quantifies the rate of completely invalid risk identifications, often referred to as false positives. Minimizing FIR is crucial for reducing spurious alerts and ensuring the efficiency of risk management efforts.

- **Expert Agreement Rate (EAR):** To quantify the level of agreement between the two expert pairs, the Expert Agreement Rate (EAR) was introduced. EAR measures the proportion of risk events for which both expert pairs reached the same validity conclusion (i.e., both pairs voted true or both pairs voted false). It is calculated as:

$$\text{EAR} = \frac{|\{r \in R \mid V_{PairA}(r) = V_{PairB}(r)\}|}{|R|} \tag{5}$$

where $V_{PairA}(r)$ and $V_{PairB}(r)$ represent the validity ratings assigned by Expert Pair A and Expert Pair B, respectively, for risk event r. EAR provides insights into the consistency and reliability of the expert validation process itself.

Additionally, to synthesize RVR, PRR, and FIR into a single composite performance measure, the **Relative Performance Index (RPI)** was introduced. RPI provides a holistic assessment of each LLM's performance, considering the trade-offs between maximizing valid risk identifications and minimizing false positives and potentially ambiguous cases. The RPI is defined as:

$$\text{RPI}(m_j) = w_1\text{RVR}(m_j) - w_2\text{PRR}(m_j) - w_3\text{FIR}(m_j) \tag{6}$$

where m_j denotes the j-th LLM model being evaluated, and w_1, w_2, and w_3 are weight factors that reflect the relative importance assigned to RVR, PRR, and FIR, respectively. For this evaluation, the weights were empirically set to $w_1 = 1.0$, $w_2 = 0.5$, and $w_3 = 1.0$. This weighting scheme emphasizes the importance of maximizing RVR (correctly identifying valid risks) and minimizing FIR (reducing false positives), while also penalizing PRR (partially validated risks) to a lesser extent. This weighting reflects a practical risk management perspective where accurate risk detection and minimizing false alarms are prioritized.

5.3 Result

This section presents the quantitative results of the multi-LLM benchmark evaluation, based on the aforementioned two-expert pairs voting framework. Two pairs of experts independently assessed each predicted risk event generated by the five LLM variants.

Tables 1 and 2 report the evaluation results obtained from Expert Pair A and Expert Pair B, respectively. These tables reflect each pair's independent evaluations and provide a detailed breakdown of the validation outcomes for each LLM. Each table includes the following metrics:

- **Count:** The total number of evaluation samples (120 risk events for each LLM).
- **V2:** The number of events where both experts in the pair voted true (fully validated by the pair).
- **V1:** The number of events where only one expert in the pair voted true (partially validated by the pair).
- **V0:** The number of events where both experts in the pair voted false (deemed invalid by the pair).
- **RVR, PRR, FIR, EAR:** The derived rates (Risk Validation Rate, Potential Risk Rate, False Identification Rate, and Expert Agreement Rate) as defined in the previous section, were calculated based on each expert pair's evaluations.

Key observations from expert pair evaluations:

- `claude3.5_sonnet`, `claude3.5_haiku`, `gpt4o`, and `gpt4o_mini` consistently demonstrate relatively high RVR values across both expert pairs, typically exceeding 0.80. This indicates that these advanced LLMs

Table 1 Evaluation results for Expert Pair A across Five LLMs

LLM	Count	V2	V1	V0	RVR	PRR	FIR	EAR
claude3.5_sonnet	120	102	14	4	0.8500	0.1167	0.0333	0.8833
claude3.5_haiku	120	100	18	2	0.8333	0.1500	0.0167	0.8500
gpt3.5_turbo	120	49	26	45	0.4083	0.2167	0.3750	0.7833
gpt4o	120	106	11	3	0.8833	0.0917	0.0250	0.9083
gpt4o_mini	120	104	13	3	0.8667	0.1083	0.0250	0.8917

Table 2 Evaluation results for Expert Pair B across Five LLMs

LLM	Count	V2	V1	V0	RVR	PRR	FIR	EAR
claude3.5_sonnet	120	107	11	2	0.8917	0.0917	0.0167	0.9083
claude3.5_haiku	120	98	19	3	0.8167	0.1583	0.0250	0.8417
gpt3.5_turbo	120	62	32	26	0.5167	0.2667	0.2167	0.7333
gpt4o	120	107	11	2	0.8917	0.0917	0.0167	0.9083
gpt4o_mini	120	106	12	2	0.8833	0.1000	0.0167	0.9000

exhibit a significantly higher rate of fully validated risk identifications compared to the baseline model.

- gpt3.5_turbo exhibits a more moderate full-validation rate (RVR), with values of 0.4083 under Expert Pair A and 0.5167 under Expert Pair B. This lower RVR, coupled with higher FIR and PRR values, suggests that gpt3.5_turbo is comparatively less effective in accurately identifying supply chain risks from unstructured text, resulting in higher rates of partial or invalid identifications.
- Both Expert Pair A and Expert Pair B exhibit overall agreement rates (EAR) at or above 0.73 for gpt3.5_turbo, and consistently above 0.84 for the other four models. These high EAR values suggest a substantial level of consistency in expert judgments, indicating that the validation process is relatively reliable and objective. The observed differences in expert judgment primarily arise around borderline or less clearly justified risk events, as reflected in the PRR values.

These metrics collectively indicate that gpt4o and gpt4o_mini tend to produce a higher proportion of fully validated risk identifications, while gpt3.5_turbo yields a larger number of cases where expert consensus is lacking or risk identification is deemed invalid. The claude3.5_sonnet and claude3.5_haiku models also demonstrate strong performance, achieving consistently high validation rates across both expert pairs, positioning them as competitive alternatives to the GPT-4o models.

While RVR is a central metric for assessing each model's reliability in risk identification, PRR captures events where expert opinions diverge, highlighting potentially ambiguous or borderline risk cases. FIR, conversely, quantifies truly spurious or irrelevant risk events. Collectively, these measures provide a comprehensive and multifaceted understanding of the models' performance under real-world risk assessment conditions, capturing both accuracy and reliability.

Aggregated performance metrics:
To derive an overall, consolidated performance metric for each LLM across both expert pairs, the mean of the four key metrics (RVR, PRR, FIR, EAR) was computed by averaging the values obtained from Expert Pair A and Expert Pair B. Additionally, the RPI was calculated for each LLM to facilitate a comprehensive comparison and inform the selection of the optimal model. Table 3 presents the mean aggregated results for each LLM, providing a unified view of model performance that integrates both experts' perspectives.

Model selection and performance ranking:
A common and practically relevant model selection criterion in SCRM is to prioritize the model with the highest RVR, reflecting the largest proportion of fully validated risk identifications, while simultaneously maintaining a low FIR to minimize false alarms. Based on this criterion, and as clearly shown in Table 3, gpt4o emerges as the top-performing model, achieving the highest RVR of 0.8875 and a remarkably low FIR of 0.0208. The second-best contenders, gpt4o_mini and claude3.5_sonnet, also exhibit strong RVR scores, but slightly trail behind gpt4o in either RVR or FIR performance. Therefore, based on the RVR and FIR metrics, gpt4o can be considered the optimal LLM choice for supply chain risk identification under this evaluation framework.

Furthermore, to encapsulate all key performance metrics (RVR, PRR, FIR) into a single evaluative measure, the RPI provides a holistic ranking of the LLMs. Based on the aggregated results presented in Table 3, the RPI values for each LLM are as follows:

Table 3 Mean Aggregated Metrics Across Expert Pairs A & B, with Relative Performance Index (RPI)

LLM	RVR	PRR	FIR	EAR	RPI
gpt4o	0.8875	0.0917	0.0208	0.9083	0.8484
gpt4o_mini	0.8750	0.1042	0.0208	0.8958	0.8334
claude3.5_sonnet	0.8708	0.1042	0.0250	0.8958	0.8250
claude3.5_haiku	0.8250	0.1542	0.0208	0.8458	0.7734
gpt3.5_turbo	0.4625	0.2417	0.2958	0.7583	0.1184

- gpt4o: RPI = 0.8484
- gpt4o_mini: RPI = 0.8334
- claude3.5_sonnet: RPI = 0.8250
- claude3.5_haiku: RPI = 0.7734
- gpt3.5_turbo: RPI = 0.1184

As evidenced by the RPI values, gpt4o achieves the highest RPI of 0.8484, indicating superior overall performance in effectively balancing risk validation and minimizing potentially valid/false identifications. gpt4o_mini and claude3.5_sonnet closely follow with RPI values of 0.8334 and 0.8250, respectively, demonstrating comparable overall performance. In contrast, gpt3.5_turbo lags significantly behind with a substantially lower RPI of 0.1184, underscoring its lower efficacy in accurate risk identification compared to the more advanced LLMs.

Considering both the RVR-FIR criterion and the holistic RPI ranking, gpt4o, gpt4o_mini, and claude3.5_haiku emerge as the top-performing models within this evaluation framework. Among these top contenders, gpt4o slightly edges out the others due to its marginally higher RVR and comparable FIR, making it the optimal choice for SCRM tasks based on the criteria established in this study. However, gpt4o_mini and claude3.5_sonnet also represent highly viable alternatives, particularly in scenarios where cost or latency considerations are paramount.

5.4 Discussion

These empirical findings from the multi-LLM benchmark confirm that newer-generation LLMs, such as GPT-4o and Claude 3.5, significantly outperform older or smaller models like GPT-3.5 turbo in the critical task of generating valid and reliable supply chain risk insights from unstructured textual data. This performance advantage is consistently observed across multiple evaluation metrics and expert validation panels.

One notable result is the consistently stronger performance of newer-generation models compared to GPT-3.5 turbo. Although the complete technical specifications of these models are not publicly disclosed, several factors likely underpin their superior ability to identify risks in unstructured supply chain news. First, newer-generation LLMs like GPT-4o are generally reported to be trained on a more extensive and diverse corpus, broadening their capacity to recognize subtle clues and context shifts. In supply chain risk scenarios–where relevant signals about disruptions can be hidden within lengthy articles or tied to specialized industry terminology–this broader knowledge base enables GPT-4o to detect more nuanced details and relationships. Second, these newer models exhibit enhanced reasoning and attention mechanisms, allowing them to link cause-and-effect patterns in news content more reliably. This improvement often results in fewer false positives and clearer interpretations of ambiguous text. By contrast, GPT-3.5 turbo–although still robust for many general tasks–tends to miss some of the implicit connections in more complex supply chain data, leading to a higher incidence of partially valid or invalid risk identifications.

Although proprietary details of each LLM's architecture remain undisclosed, the empirical results strongly suggest that advanced training processes and refined attention mechanisms contribute to higher RVR and lower FIR for these newer models. By leveraging deeper context windows, improved instruction-following, and broad pretraining on diverse data, LLMs such as GPT-4o and Claude 3.5 show better generalization when classifying complex risks. Consequently, companies aiming to automate risk detection may benefit more from these newer-generation LLMs–provided they also institute safeguards such as human validation and ongoing fact-checking to mitigate residual errors and biases.

While these metrics (RVR, PRR, and FIR) provide an overview of the system's precision and reliability, our observations also revealed occasional inaccuracies or "hallucinations" in the model outputs–particularly in borderline cases where context was sparse. This underscores the importance of human validation steps to ensure that AI-driven risk assessments remain accurate, especially when dealing with sensitive or high-stakes supply chain decisions.

For organizations with the necessary budgets and high-stakes supply chain decisions, GPT-4o stands out as the most reliable and accurate model, offering the highest RVR and lowest FIR in our tests. However, for those with more limited resources or heightened latency constraints, smaller and more cost-effective options like GPT-4o mini and Claude 3.5 Sonnet still offer strong performance, striking a useful balance between capability and resource efficiency. Ultimately,

the most suitable LLM choice will hinge on an organization's specific operational context, budget considerations, and performance requirements.

5.5 Strengths of LLM-RI

Building upon the empirical evidence, this section consolidates the overarching strengths of the LLM-RI subframework with the key design decisions that enable these strengths. Taken together, these elements underscore the robust performance of LLM-RI in proactively identifying a wide spectrum of supply chain risks from large volumes of unstructured data.

5.5.1 High accuracy and reliability in real-world data analysis

A primary strength of the LLM-RI approach is its demonstrably high degree of accuracy in identifying relevant risk events within real-world datasets derived from news articles. The consistently high Risk Validation Rates (RVR) observed across both single-LLM and multi-LLM evaluations, coupled with expert validation, affirm the system's precision in detecting genuine supply chain risks. This robust performance is maintained even when employing different state-of-the-art LLMs, highlighting the generalizability and reliability of the methodology. This accuracy and reliability are crucial for building confidence in the system's outputs and enabling informed decision making in risk management.

5.5.2 Holistic and novel risk capture beyond predefined dictionaries

A significant advantage of the LLM-RI methodology lies in its inherent capacity to discern and interpret textual references to novel, emergent, or previously unknown risks. Unlike conventional rule-based or keyword-driven approaches that rely on predefined dictionaries of known risks, LLM-RI leverages the nuanced language understanding capabilities of modern LLMs to identify risks that may not be explicitly anticipated or captured by traditional methods. This ability to detect subtle signals of disruption and identify unforeseen risk factors is a critical asset in the dynamic and complex landscape of contemporary supply chains, enhancing overall risk identification coverage and proactive risk management capabilities.

5.5.3 Scalability and automation for efficient risk monitoring

The system's design inherently supports high scalability and automation, enabling efficient and continuous risk monitoring across extensive SCNs. LLM-RI is engineered to autonomously process hundreds of news articles daily, covering dozens or even hundreds of suppliers, with minimal manual supervision. This scalability and automation are essential for effectively managing the vast volume of unstructured data relevant to supply chain risk and for ensuring timely and resource-efficient risk monitoring. The automated nature of LLM-RI facilitates rapid analysis and continuous surveillance, enabling near-real-time detection of emerging risks and proactive intervention.

6 Conclusion and future work

In this paper, we extended our prior efforts on the LARD-SC framework by integrating and benchmarking a diverse set of state-of-the-art LLMs for supply chain risk identification. Our work demonstrated that the incorporation of advanced AI techniques into traditional risk management practices can significantly enhance the accuracy and efficiency of risk detection, classification, and visualization. Through rigorous evaluation using a standardized prompt-based approach and expert assessments, our results highlighted that GPT-4o variants, in particular, exhibit superior capabilities in identifying fully validated risk events while minimizing false positives compared to earlier models. The comprehensive quantitative analysis using metrics such as RVR, PRR, FIR, and RPI provided robust empirical evidence to guide practitioners in selecting the most suitable LLMs for their SCRM applications.

Beyond illustrating the technical advances achievable with current LLMs, this study has also underscored several practical considerations. While our evaluations confirmed the promise of enhanced risk identification, they also revealed challenges inherent in LLM deployments, such as occasional inaccuracies, dependency on prompt quality, and the need for fine-grained expert validations. These challenges emphasize that, despite the significant progress made, effective integration of LLM-driven analytics into decision support systems must balance technological sophistication with careful human oversight.

Building on the insights gained from this study, several avenues for future research and development are apparent:

1. **Incorporating statistical testing and correlation analyses:** Although this study reports various performance metrics for the evaluated LLMs, further work could involve implementing formal significance testing (e.g., t-tests or ANOVA) to rigorously evaluate the robustness of observed performance differences. Additionally, conducting correlation analyses to examine how factors such as news article characteristics or model configurations influence accuracy would provide deeper insights into why one model may outperform another in particular

scenarios. These quantitative enhancements would build on the existing expert validation framework, ultimately strengthening both the reliability and interpretability of the system for more informed LLM-driven risk identification within complex supply chain contexts.

2. **Extending multimodal data sources:** Global supply chain risks are not solely discernible through textual analysis; they often manifest through visual or auditory cues embedded in images, videos, and other non-textual media. Future research should concentrate on integrating advanced multimodal data processing techniques into the LARD-SC framework. By incorporating state-of-the-art vision-language models and audio analysis algorithms, the system can extract supplementary risk indicators from diverse media formats. This enhancement would facilitate a more comprehensive evaluation of potential disruptions, enabling risk managers to obtain richer, multi-dimensional insights that could significantly improve the detection and mitigation of supply chain vulnerabilities.
3. **Mitigating hallucinations and bias in LLM outputs:** While the application of LLMs has significantly advanced automated risk detection, these models can sometimes generate outputs that are either unverified or influenced by inherent biases present in their training data. To address these concerns, future work should investigate the integration of additional fact-checking layers and cross-referencing mechanisms to validate extracted risk signals. Moreover, developing robust bias detection and correction protocols is essential to ensure that the risk assessments are both accurate and equitable, thereby reinforcing RM confidence in the automated system.
4. **Enhancing domain-specific customization:** Supply chains across different sectors, such as pharmaceuticals, aerospace, and food production, encounter unique challenges that necessitate tailored risk management approaches. Future research should explore the customisation of LLM prompts and risk taxonomies to better align with sector-specific characteristics. Furthermore, by incorporating domain-specific fine-tuning and integrating human-in-the-loop feedback, the framework can achieve greater granularity and contextual relevance in risk identification, assessment and classification, ultimately enhancing the precision of automated risk management.
5. **Leveraging next-generation LLMs for enhanced risk detection and decision support:** Recent advancements in next-generation LLMs offer substantial promise for further refining the LARD-SC framework. These emerging models exhibit enhanced contextual reasoning, improved multi-modal integration, and superior dynamic learning capabilities, all of which are critical for maintaining accurate and timely risk identifications and assessments. Future research should focus on integrating these next-generation LLMs to bolster risk detection accuracy, reduce false positives, and provide richer, explainable insights. Additionally, adaptive learning mechanisms inherent in these advanced models can facilitate real-time updates, ensuring that the risk management system remains effective in the face of evolving global threats.
6. **Ethical and regulatory considerations:** As reliance on AI-driven risk management tools increases, addressing issues related to data privacy, fairness, and accountability becomes critical. Future research should examine these ethical and regulatory dimensions in greater depth, developing guidelines and best practices to ensure responsible and transparent AI integration in SCRM.

In summary, this study has laid a strong foundation for the next generation of SCRM systems by illustrating the transformative potential of LLMs within the LARD-SC framework. As AI technologies continue to evolve, ongoing research and development will be essential to fully realize their benefits while mitigating their limitations, paving the way for more resilient and adaptive service-oriented SCRM practices in the future.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions The first author acknowledges the financial support received from The University of New South Wales for this work.

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cambridge Centre for Risk Studies. Cambridge Taxonomy of Business Risks. Cambridge Centre for Risk Studies; 2019
2. Zhao M, Hussain O, Zhang Y, Saberi M (2024) Optimizing supply chain risk management: An integrated framework leveraging large language models. In: 2024 IEEE Conference on Artificial Intelligence (CAI). IEEE p. 1057–1062

3. Zhao M, Hussain O, Zhang Y, Saberi M, Leshob A (2024) Enhancing Supply Chain Risk Management with Large Language Models: Software Prototyping and Interactive Visualization. In: 2024 IEEE International Conference on e-Business Engineering (ICEBE). IEEE p. 284–291
4. Christopher M (2016) Logistics and supply chain management: logistics & supply chain management. Pearson UK;
5. Ho W, Zheng T, Yildiz H, Talluri S (2015) Supply chain risk management: a literature review. Int J Prod Res 53(16):5031–5069
6. Heckmann I, Comes T, Nickel S (2015) A critical review on supply chain risk-definition, measure and modeling. Omega 52:119–132
7. OpenAI.: OpenAI [Web Page]. Available from: https://openai.com/
8. Anthropic.: Anthropic [Web Page]. Available from: https://www.anthropic.com/
9. Bubeck S, Chadrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E et al. Sparks of artificial general intelligence: Early experiments with gpt-4. ArXiv
10. Kleindorfer PR, Saad GH (2005) Managing disruption risks in supply chains. Prod Oper Manag 14(1):53–68
11. Tang CS (2006) Perspectives in supply chain risk management. Int J Prod Econ 103(2):451–488
12. Ivanov D (2022) Viable supply chain model: integrating agility, resilience and sustainability perspectives-lessons from and thinking beyond the covid-19 pandemic. Ann Oper Res 319(1):1411–1431
13. Giannakis M, Papadopoulos T (2016) Supply chain sustainability: a risk management approach. Int J Prod Econ 171:455–470
14. Choi TM, Chan HK, Yue X (2016) Recent development in big data analytics for business operations and risk management. IEEE transactions on cybernetics 47(1):81–92
15. Bai C, Dallasega P, Orzes G, Sarkis J (2020) Industry 4.0 technologies assessment: A sustainability perspective. International journal of production economics. 229 107776
16. Waller MA, Fawcett SE. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. Wiley Online Library
17. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al (2017) Attention is all you need. Advances in neural information processing systems. 30
18. Mahmud D, Hajmohamed H, Almentheri S, Alqaydi S, Aldhaheri L, Khalil RA, et al (2025) Integrating LLMs with ITS: Recent Advances, Potentials, Challenges, and Future Directions. IEEE Transactions on Intelligent Transportation Systems
19. Cheng ZQ, Dong Y, Shi A, Liu W, Hu Y, O'Connor J, et al (2024) Shield: Llm-driven schema induction for predictive analytics in ev battery supply chain disruptions. arXiv preprint arXiv:2408.05357
20. Gaurav M (2023) Building Supply Chain Resilience Using Artificial Intelligence in Risk Management Systems. In: Smart Services Summit. Springer; p. 55–67
21. Jackson I, Ivanov D, Dolgui A, Namdar J (2024) Generative artificial intelligence in supply chain and operations management: a capability-based framework for analysis and implementation. Int J Prod Res 62(17):6120–6145
22. Kühl L, Wiethölter J, Dircksen M (2024) Enhancing Supply Chain Risk Identification: Analyzing the Impact of LLM Parameters for precise Classification. In: Symposium on Logistics; p. 197
23. Floridi L, Cowls J (2022) A unified framework of five principles for AI in society. Applications in architecture and urban design, Machine learning and the city, pp 535–545
24. Xu M, David JM, Kim SH et al (2018) The fourth industrial revolution: opportunities and challenges. International journal of financial research 9(2):90–95
25. OpenAI.: GPT-3.5 Turbo [Web Page]. Available from: https://platform.openai.com/docs/models#gpt-3-5-turbo
26. OpenAI.: GPT-4o [Web Page]. Available from: https://openai.com/index/hello-gpt-4o/
27. OpenAI.: GPT-4o mini [Web Page]. Available from: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/
28. Anthropic.: Claude 3.5 Sonnet [Web Page]. Available from: https://www.anthropic.com/claude/sonnet
29. Anthropic.: Claude 3.5 Haiku [Web Page]. Available from: https://www.anthropic.com/claude/haiku