SN Computer Science (2026) 7:645
https://doi.org/10.1007/s42979-026-05213-z

ORIGINAL RESEARCH

Large Language Models for Risk Detection in E-commerce: Reliability, Semantic Alignment, and Managerial Insights

Laleh Davoodi[1] · Filip Ginter[1] · Sima Salimi[2] · Harri Lorentz[1]

Received: 19 December 2025 / Accepted: 30 June 2026

Abstract
The increasing complexity of global e-commerce supply chains underscores the need for automated, context-aware risk monitoring systems capable of interpreting large volumes of unstructured information. Although Large Language Models (LLMs) have shown strong performance across natural language processing tasks, their application to real-world supply chain risk detection remains limited. This study presents a novel, manually annotated dataset of 121 business news articles related to five major steel companies, using the Cambridge Risk Taxonomy. Leveraging this dataset, we evaluate two state-of-the-art LLMs in a multi-label risk classification task using few-shot prompting. The results demonstrate that LLMs can approximate human annotation, though challenges persist in detecting domain-specific risks such as Geopolitical threats and in avoiding label overgeneration. Beyond classification, we further assess the capacity of LLMs to generate managerial risk summaries. We show that summaries derived from model-predicted risks exhibit strong semantic alignment to summaries generated from human annotations, highlighting the potential of LLMs to support executive-level risk interpretation. Overall, this study contributes the first publicly available dataset of fine-grained, hierarchical risk annotations in an e-commerce supply chain context and provides empirical evidence on the opportunities and limitations of LLMs for both analytical and narrative forms of automated risk assessment.

Keywords E-commerce · Business news analysis · Risk management · Machine learning · LLMs

Introduction

The digital transformation of global markets has significantly reshaped business operations, with Europe's e-commerce infrastructure emerging as a central component of the digital economy. As presented in Fig. 1, in 2025, Europe is projected to generate approximately $ 707.9 billion in retail e-commerce revenue, placing it third globally after Asia and the Americas [30].

E-commerce has become a vital component of Europe's economy, generating billions in revenue and accelerating the shift away from traditional retail practices. This growth is driven by technological advancements, rising internet penetration, and evolving consumer expectations centered on convenience, personalization, and sustainability [12].

As complex systems integrating logistics, payments, and customer service across interconnected digital infrastructures, e-commerce platforms involve multiple stakeholders and therefore face diverse risks, including fraud, data breaches, platform outages, regulatory changes, reputational challenges, and commodity price fluctuations [17]. These risks intersect with the longstanding vulnerability of supply chains to disruptions caused by natural disasters, economic or political crises, and unexpected accidents. Modern interconnected networks amplify the effects of such incidents, making even minor disturbances propagate rapidly across supply chain nodes. This heightened interconnectedness underscores the need for proactive Supply Chain Risk

✉ Laleh Davoodi
ladavo@utu.fi

Filip Ginter
figint@utu.fi

Sima Salimi
sima.salimi@abo.fi

Harri Lorentz
harri.lorentz@utu.fi

[1] University of Turku, Turku FI-20014Turun yliopisto, Finland

[2] Åbo Akademi University, Tuomiokirkontori 3, 20500 Turku, Finland

Published online: 22 July 2026

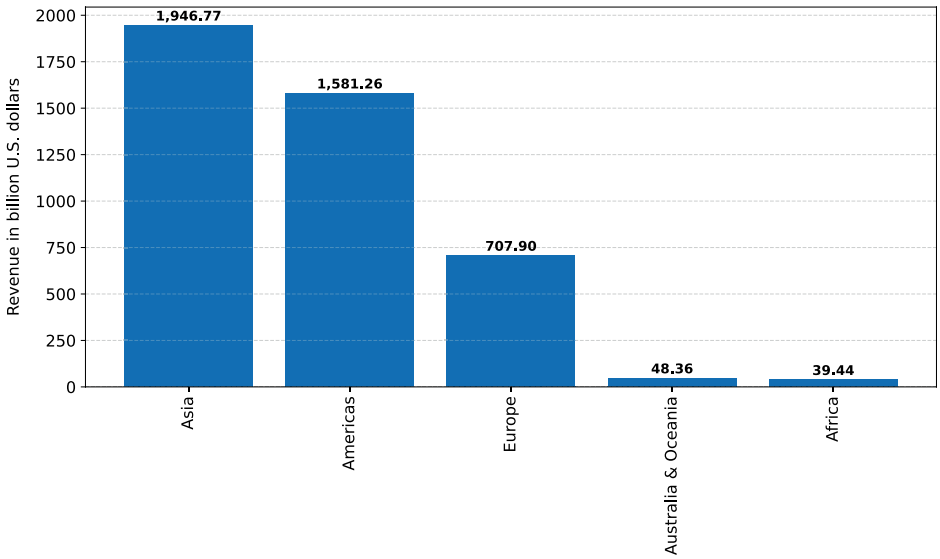

Fig. 1 Total retail e-commerce revenue worldwide in 2025, by region (in billion U.S. dollars). Source: Statista Digital Market Insights, June 2025

Management (SCRM) within contemporary Supply Chain Management (SCM) frameworks [74].

SCRM involves identifying and mitigating risks through coordinated strategies across supply chain partners and has become increasingly essential due to the frequency and long-term performance impacts of supply chain disruptions [73]. As Dutta et al. [36] demonstrate, the scale and speed of digitally enabled commerce, whether B2C or B2B, amplify the consequences of cybersecurity threats, logistics bottlenecks, and platform failures, making effective SCRM crucial for reliability, trust, and competitiveness.

Despite increasing exposure to operational and strategic risks within digitally enabled e-commerce ecosystems, the systematic use of advanced AI tools, particularly Large Language Models (LLMs), for automated risk detection remains underdeveloped. Moreover, the literature lacks any publicly available, manually annotated dataset for identifying supply-chain-relevant risks in real-world business news. This study addresses this gap by introducing a novel, expert-annotated dataset of risk-labeled news articles covering five major steel producers: ArcelorMittal, Tata Steel, POSCO, ThyssenKrupp, and NLMK. Although these firms are not conventionally situated within the e-commerce domain, their procurement, logistics coordination, and market access processes are increasingly mediated through digital platforms, rendering them integral participants in contemporary e-commerce–enabled supply networks.

The European steel industry is a cornerstone of the EU economy, contributing over € 215 billion in turnover, supporting millions of jobs across more than 500 production sites, and underpinning key manufacturing, construction, and sustainable development activities through its innovative, versatile, and fully recyclable material base [75]. As Ramadhan et al. [48] demonstrate, SCRM is especially critical in the steel industry, which is characterized by chronic vulnerabilities such as raw material scarcity, price volatility, production capacity constraints, logistical disruptions, and evolving regulatory pressures. These risks are further exacerbated by the sector's dependence on imported inputs and its sensitivity to global economic and geopolitical fluctuations. Against this backdrop, the dataset introduced in this study provides a valuable resource for assessing the capability of LLMs to detect and classify business risks embedded in unstructured news texts.

Beyond label-level evaluation, we argue that managerial usefulness depends on whether an automated system conveys a faithful "risk story" about a company, not merely whether each fine-grained label matches a human annotation. To that end, we augment the standard classification analysis with a narrative evaluation: for each article, we generate manager-oriented risk summaries from (i) human annotations and (ii) model predictions, then assess their semantic alignment and human-centric quality. This dual perspective, including labels and summaries, allows us to test whether LLMs can both recognize risks and communicate them in a way that supports executive decision-making.

Concretely, our contributions are threefold. First, we release a domain-specific, manually annotated news corpus with hierarchical risk labels suitable for multi-label modeling and replication. Second, we implement a retrieval-augmented, few-shot prompting pipeline and provide a head-to-head comparison of GPT-4o and LLaMA−3.3–70B on real-world, long-form news. Third, we introduce a summary-level reliability assessment in which model-generated narratives are compared to annotation-based narratives, showing that even when some labels are misassigned, the resulting summaries can remain semantically aligned and managerially coherent.

From a digital services perspective, such LLM-based risk monitoring systems function as pervasive, continuously operating decision-support services that translate large volumes of unstructured information into accessible, human-interpretable risk insights, thereby supporting transparent, timely, and more sustainable decision-making across e-commerce ecosystems. Therefore, our goals are organized into the following research questions:

RQ1: How can unstructured business news be systematically annotated with risk information that reflects e-commerce and industrial realities using a comprehensive taxonomy?

RQ2: To what extent can contemporary LLMs detect and classify business risks from long-form, real-world news under few-shot prompting?

RQ3: Do model-generated summaries preserve the semantic content and communicative quality of human-aligned risk narratives, thereby supporting managerial interpretation even when some fine-grained labels differ?

By combining label-level metrics with narrative-level analysis, the paper advances beyond conventional accuracy reporting to evaluate whether LLMs can deliver executive-ready risk briefings, an outcome crucial to the practical deployment of LLMs in e-commerce and supply-chain contexts.

Literature Review

As digitally integrated supply chains become more complex, particularly in sectors where procurement, logistics coordination, and market access increasingly depend on e-commerce platforms, effective risk management has become essential. Although LLMs show promise for real-time risk detection, existing studies mostly rely on synthetic or narrowly scoped data. This review identifies a gap in applying LLMs to real-world, manually annotated business news, especially in industrial sectors such as steel, where supply chains are deeply interconnected with global e-commerce-enabled procurement and logistics networks.

Supply Chain Risk Management

The theory of Supply Chain Risk Management (SCRM) was introduced in the late 1990s and early 2000s. It was developed progressively through multiple scholarly contributions rather than a single foundational work. Tang [8] outlined key dimensions of SCRM, including risk classification, mitigation strategies, and the integration of risk awareness into strategic decision-making processes. Around the same time, Simchi-Levi et al. [35] also advanced the theory by introducing practical frameworks and analytical tools that brought SCRM into mainstream academic and industry conversations.

Subsequent research underscored the role of SCRM in sustaining market competitiveness and ensuring operational resilience. Hillman and Keltz [37] highlight that firms are increasingly recognizing the importance of managing risks within the supply network to protect their market position and financial viability. While early SCRM frameworks primarily focused on operational disruptions such as supplier failures, transportation breakdowns, or demand fluctuations [7, 8], subsequent studies have expanded the concept to include strategic, financial, and governance risks [51].

As businesses frequently encounter diverse risks, effective classification is essential. Risks are often grouped as strategic, tactical, or operational, depending on their impact on long-term goals, internal processes, or daily operations. Another approach distinguishes between internal and external risks, with internal risks being more controllable, while external risks, such as political or economic disruptions, are harder to manage. Identifying and categorizing these risks, particularly those arising from unstructured information, is crucial for developing effective risk management strategies [33].

In recent years, however, the rapid digitalization of business processes and the emergence of e-commerce have transformed the scope of risk management. These developments introduce new vulnerabilities and interdependencies that extend beyond traditional supply-chain disruptions and require fresh analytical approaches, which are discussed in the next section.

Supply Chain Risk Management in E-commerce

With the advent of technology, especially internet-based commerce and payment, E-commerce has emerged as a disruptive channel for many companies that reshape consumer behavior, supply chains, and competitive dynamics [38]. E-commerce refers to the digital buying and selling of goods and services, reshaping markets and consumer behavior through models like B2B and B2C. While it offers convenience and competitiveness, it also introduces new risks that can threaten business operations, trust, and long-term success [15].

The risk landscape in e-commerce is multifaceted. E-business initiatives introduce a range of risks that differ significantly from traditional business environments due to their dynamic, fast-evolving nature. These risks span strategic, operational, legal, financial, and technological domains, with particular emphasis on challenges such as security breaches, system dependencies, intellectual property violations, and reputational threats [27].

Moreover, Dutta and his colleagues discuss new dimensions in managing risk for e-commerce supply chains. Their study focuses on various risks associated with inventory management and strategies to mitigate these risks. Key risks identified include contract compliance, employee and third-party fraud, demand fluctuations, and challenges related to reverse logistics. Several mitigating strategies are proposed, such as pre-purchase stock, Just in Time inventory, drop shipping, and a hybrid model [36].

SCRM frameworks must now account for digital dependencies, cybersecurity, and platform-based interactions [52]. For instance, cybersecurity threats, such as social engineering, malware, and denial-of-service attacks, remain prominent challenges with direct implications for platform stability and user safety. Furthermore, the digital nature of e-commerce exposes firms to complex Financial, Regulatory, Geopolitical, and Governance risks, often driven by external macro-environmental changes. These risks are difficult to detect with traditional rule-based systems due to their gradual and interconnected nature. While SCRM helps address operational and logistical disruptions, it covers only part of the broader risk landscape relevant to e-commerce [8].

It also considers strategic decisions, such as sourcing models, inventory structures, and information-sharing practices, that influence resilience. In this context, prior studies have categorized risks as arising from supply, demand, process, political, and environmental domains [2, 10].

However, many of the risks relevant to e-commerce businesses today, especially those operating across borders, extend beyond supply chain disruptions. For instance, Geopolitical risks refer to the potential emergence, actual occurrence, or escalation of adverse events, such as wars, terrorism, or tensions among states and political actors that threaten the stability and peaceful functioning of international relations [31].

It comprises a wide array of interconnected factors, including the threat of military conflict, economic competition, the rise of new, often non-state actors, climate-related risks, localized or global military escalations, and the involvement of emerging players in already volatile situations, such as terrorist or para-state groups. As a result, Geopolitical risk is complex and multifaceted. It has become an increasingly influential force shaping economic performance, financial markets, and the behavior of economic agents across multiple countries [32].

On the other hand, Governance risks include strategic missteps, regulatory non-compliance, and reputational crises, while Financial risks may emerge from commodity price volatility, investor outlook, or macroeconomic shocks [22]. These risk categories, drawn from comprehensive taxonomies such as the Cambridge Risk Taxonomy [22], offer a more holistic view of the threats facing modern e-commerce businesses. This reinforces the need for risk detection systems that are capable of capturing subtle and interrelated risk signals across diverse domains, particularly in high-stakes industrial contexts such as steel and infrastructure-linked e-commerce.

Supply Chain Risk in the Steel Industry

The steel industry supply chain is one of the most complex and globally integrated industrial systems, including multiple stages such as raw material extraction, ironmaking, steel production, processing, and distribution to end-use sectors. Managing this multi-stage process requires synchronization of production, inventory, and organizational logistics levels [43].

Due to its dependence on commodities such as iron ore, coal, or gas, and energy, as well as large-scale infrastructure and transportation systems, the steel supply chain is exposed to economic volatility, geopolitical instability, and environmental regulations [44].

Xiong and Helo's research [42] emphasises that fluctuating demand, raw-material supply bottlenecks, and uncertain pricing create significant complexity for the steel supply chain. Numerous studies have analyzed various controllable risks and their interdependencies within the supply chain. Venkatesh et al. [40] categorizes these risks based on their driving and dependence, revealing that globalization, labor issues, and resource security are significant factors in business.

However, there is no universal specific classification framework that all businesses use. Risk categories often vary depending on the nature and structure of the business. Consequently, various studies that have examined and prioritized supply chain risk within specific businesses. For instance, Üstündağ et al. [41] found the most critical risk group, followed by customer risks, supplier risks, transportation risks, environmental risks, and security risks in the steel business. Table 1 summarizes selected studies that have explored the type and prioritization of risks in steel supply chains. In addition, the identified risks in each article are aligned with categories defined in the Cambridge Risk Taxonomy [22]. This comparative approach clarifies which risk types are most frequently emphasized in the literature.

As presented in Table 1, the results of eight studies on steel supply chain risk demonstrate that Governance risks are the most emphasized category across the literature. This reflects the center of supply chain risk across internal management, operational control, and compliance mechanisms. Financial risks in the second rank highlight the steel industry's high exposure to commodity price fluctuations.

Table 1 Summary of research on steel supply chain risks and adoption, mapped to cambridge taxonomy categories

	Year	Article summary	Cambridge category	Link
1	2008	*Operational* (production, logistics coordination); *Financial* (price volatility, energy cost); *Market* (cyclical demand changes); *Supply* (raw material dependency); *Logistics* (transportation and infrastructure)	Governance financial strategic	[42]
2	2022	*Supplier* (delay in material delivery, supplier insolvency, quality inconsistency); *Customer* (demand uncertainty, changes in order quantity, payment delays); *Transportation* (logistics delays, port strikes, damage during transit); *Environmental* (political uncertainty, natural disasters, economic crisis); *Security* (cyber attacks, natural disaster); *Process* (insufficient or excess inventory)	Governance strategic environmental technology	[41]
3	2023	*Environmental* (pollution, emissions, resource use); *Financial* (price and cost volatility); *Technological* (innovation lag, inefficiency); *Governance* (weak policy coordination); *Geopolitical* (trade instability)	Environmental governance strategic technology geopolitical	[45]
4	2024	*Financial* (price volatility, exchange rate); *Governance* (internal sourcing and management control); *Technological* (outdated technology, automation gap)	Governance financial technology	[46]
5	2024	*Societal* (stakeholder pressure, CSR); *Financial* (budget overruns, price volatility); *Environmental* (resource depletion, pollution); *Technological* (digitalization gap)	Financial governance environmental technological	[47]
6	2024	*Human Resource Risk* (lack of skilled workers); *Technological* (digital integration); *Financial* (cash flow); *Logistics* (transport delays); *Supply* (supplier reliability, quality)	Governance financial technological societal	[48]
7	2025	*Societal* (stakeholder pressure, CSR); *Financial* (budget overruns, price volatility); *Environmental* (resource depletion, pollution); *Technological* (digitalization gap)	Financial governance environmental technological	[49]
8	2025	*Supply* (import complexity); *Logistics* (Port congestion); *Financial* (raw material price volatility); *Geopolitical* (Trade conflicts); *Environmental* (Decarbonization pressure); *Managerial* (Limited scenario planning)	Governance geopolitical financial environmental	[50]

Also, Environmental and Technological risks have grown in recent years that driven by sustainability imperatives and pushing technology. Societal and Geopolitical risks represent emerging external pressures that are becoming important with global steel interconnected markets and policy-driven. In addition, the results also indicate that the Cambridge Risk Taxonomy provides a multidimensional design to categorize all risk types identified in the reviewed literature. Each risk type discussed in the eight analyzed studies can be mapped to one or more of the Cambridge categories. This alignment confirms that the Cambridge taxonomy is not only appropriate for traditional supply chain risk classifications (such as supplier, logistics, and market risks) but also extends to emerging dimensions, including governance and technology.

AI in Risk Assessment

We are living in a technology-driven era where Artificial Intelligence (AI) has become deeply embedded in daily life, from smart devices to automated systems. In the context of e-commerce, AI is transforming business operations, enhancing customer interactions through chatbots, improving product recommendations, enabling large-scale data processing, and personalizing user experiences [16].

At the core of this transformation lie LLMs. LLMs have demonstrated exceptional capabilities across a wide range of natural language processing (NLP) tasks and related domains. Their success has spurred extensive research on architectural innovations, improved training strategies, extended context handling, fine-tuning, multimodal learning, benchmarking, and efficiency optimization [57].

The Generative Pre-trained Transformer (GPT) series, developed by OpenAI, represents a major advancement in NLP by enabling systems such as ChatGPT to understand and generate human-like text with high accuracy. Built on the transformer architecture, GPT models have achieved state-of-the-art results in a variety of tasks, leading to widespread adoption in both academia and industry [58]. For instance, among the most influential, OpenAI's GPT-3 could generate text that is nearly indistinguishable from human writing, marking a milestone in machine language understanding [56].

The LLaMA (Large Language Model Meta AI) family, developed by Meta, represents some of the most competitive openly available large language models. Notably, LLaMA-13B surpasses GPT-3 while being over ten times smaller, and LLaMA-65B performs comparably to Chinchilla-70B and PaLM-540B. Unlike previous models, LLaMA demonstrates that competitive performance can be achieved using only publicly available datasets [59]. In LLMs, the number of parameters reflects their complexity and learning capacity; larger models can capture richer linguistic context and subtleties, leading to more coherent and context-aware text generation [60].

LLMs have transformed text classification by reducing the need for extensive preprocessing, domain-specific expertise, and large manually labeled datasets. Recent studies show that LLM-based classifiers can outperform traditional machine-learning and neural models across tasks

such as sentiment analysis, spam detection, and multi-label classification, especially when enhanced with few-shot prompting or fine-tuning [65–67].

These advances are directly relevant for SCRM, where risk identification fundamentally relies on classifying unstructured textual sources such as news articles [68]. As SCRM increasingly depends on the timely interpretation of large volumes of heterogeneous information, LLM-based text classification provides a scalable mechanism for detecting emerging risks, filtering relevant signals, and mapping textual evidence to structured taxonomies [69].

Recent work highlights how LLMs can be integrated into SCRM pipelines through zero-shot and few-shot classification, fine-tuning, and retrieval-augmented generation (RAG), enabling automated extraction of both analytical (risk labels) and narrative (managerial summaries) insights. These approaches have been applied across diverse industrial contexts. For instance, Shahsavari et al. developed CERIA, a hybrid model combining LLMs and Bayesian Networks to detect and assess risks from news data through event detection and probabilistic reasoning [18]. Cheng et al. proposed SHIELD, a two-stage system for the EV battery supply chain, integrating GPT-based schema learning with fine-tuned RoBERTa and Graph Convolutional Networks, thereby improving interpretability and accuracy [18]. Zhu and Vuppalapati showed that RAG enhances forecasting, supplier evaluation, and risk mitigation by combining retrieval and generation mechanisms, outperforming fine-tuned models in contextual reasoning [23]. Similarly, Quan and Liu introduced InvAgent, a multi-agent inventory management system leveraging zero-shot LLMs and Chain-of-Thought reasoning to dynamically optimize stock levels without pre-training [24].

In financial risk contexts, Teixeira et al. proposed Labeled Guide Prompting (LGP), which improves LLM reliability in credit risk analysis by combining annotated few-shot examples with structured Bayesian guidance, achieving human-preferred outputs in up to 90% of cases [25]. Koli et al. presented an AI-driven Insider Risk Management (IRM) system that integrates behavioral analytics, adaptive risk scoring via autoencoders, and real-time policy enforcement, improving detection accuracy by 30% while significantly reducing false positives [26].

Collectively, these studies demonstrate the expanding role of LLMs in enabling proactive, interpretable, and adaptive risk management across supply chains, as summarized in Table 2. Their integration with structured reasoning approaches, such as Bayesian inference, schema learning, and graph analytics, has further enhanced contextual understanding and decision support. However, most existing studies rely heavily on synthetic, narrow, or domain-specific datasets, leaving substantial opportunity for advancing

Table 2 Overview of recent LLM applications in supply chain research

Model	Methodology	Domain	Findings	Dataset	Cite
OpenAI (GPT-4, GPT-4o, GPT-4-Turbo)	Prompt Engineering (Zero-shot Prompting) & CoT	Inventory Management	Demonstrates robustness and adaptability in dynamic supply chains	–	[24]
Gemini 1.5 Flash & Gemini 1.5 Pro	Prompt Engineering (Sequential Prompting and few-shot Prompting)	Inventory Management (Sequential Supply Chain Optimization)	Model size and tool reliability do not guarantee high performance	–	[63]
OpenAI (GPT−3.5-Turbo)	Prompt engineering (few-shot approach)	Risk Identification in Supply Chains	Strong event impact analysis in supply chain scenarios	News data	[18]
Unknown Model	Prompt engineering (few-shot approach)	Supply Chain Transparency	Enhances visibility in network structures through LLM-powered analysis	Bloomberg SPLC, Internet Text	[19]
Unknown Model	Prompt engineering (few-shot approach) with RAG	Supply Chain Optimization	Improves decision-making by integrating external knowledge sources	ERP, CRM, TMS, Market Reports	[23]
OpenAI (GPT−3.5-Turbo-1106)	Prompt engineering (few-shot approach)	Risk Identification & Assessment	Significantly improves risk detection and management efficiency	Google News Articles	[21]
Unknown Model	Semantic search and NLP-based text embeddings	Event Identification in Supply Chains	Automates risk detection using large-scale data	Google News Articles	[20]

real-world e-commerce risk analytics. Moreover, although LLM-based approaches increasingly achieve strong algorithmic accuracy, far less is known about their interpretive reliability, especially in producing faithful and decision-relevant narratives from unstructured business news. This concern is amplified by known limitations of LLMs, including hallucination [70], overgeneralization [72], and sensitivity to prompt phrasing [71], all of which can undermine trustworthiness in high-stakes supply chain settings.

These issues are particularly problematic in decisive supply chain settings, where inaccurate or fabricated risk signals may misguide decision-makers. Therefore, evaluating LLM outputs solely through label-level accuracy is insufficient. There is a need for assessment approaches that examine whether LLMs can produce faithful, contextually grounded, and decision-relevant narratives. To address these gaps, the present study introduces a dual evaluation framework, assessing both label-level classification and summary-level narrative fidelity, to advance the development of interpretable, human-aligned, and managerially actionable risk analytics.

Methodology

To explore the potential of LLMs for automated risk detection in real-world e-commerce, we created a novel, manually annotated dataset of 121 news articles about five major steel companies engaged in digital commerce, using the Cambridge Risk Taxonomy. The annotation process emphasized consistency and contextual accuracy, and two state-of-the-art LLMs were evaluated using few-shot prompting with semantically similar examples. This section outlines the dataset design, taxonomy adjustments, and model evaluation pipeline for the multilabel classification task.

Data Gathering and Data Annotation

The news articles used in this study were obtained through an institutional subscription to the CorpusData database,[1] which provides access to a large collection of international news sources for research purposes. The texts provided by CorpusData are already processed and anonymized; therefore, the anonymization was performed by the data provider, not by the authors.

From this source, a dataset of 121 news articles was curated covering five steel companies operating in the e-commerce sector: POSCO, ArcelorMittal, Tata Steel, NLMK Group, and ThyssenKrupp. The temporal scope of the search was guided by the approximate period when each company initiated or significantly expanded its e-commerce activities, ranging from 2001 (POSCO) to 2017 (ArcelorMittal). These dates were used only to establish the lower bound for identifying relevant news related to e-commerce developments. The final dataset consists of articles published between 2010 and 2023, depending on relevance and data availability.

The dataset size reflects the length of the articles (average 736 words), which rendered manual annotation time-intensive (1–2 h per article). This scale is also consistent with recent LLM-based supply chain risk studies [34, 64], which likewise rely on manually curated datasets of comparable size. Beyond creating a labeled dataset, the study aimed to explore the annotation process and evaluate how LLMs handle the complex, subjective task of risk identification. For this purpose, two annotators independently labeled the articles using the Cambridge Risk Taxonomy across three batches.

The Cambridge Risk Taxonomy defines six key business risk classes: *Financial, Geopolitical, Technology, Environmental, Social, and Governance*. *Financial risks* involve macroeconomic changes like interest rates; *Geopolitical risks* arise from political instability and policy shifts; *Technology risks* include cyber threats and digital disruptions; *Environmental risks* relate to natural hazards and sustainability; *Social risks* stem from consumer behavior, public health, and reputational issues; and *Governance risks* cover compliance, litigation, and leadership failures, all of which can significantly affect corporate performance [22].

After completing individual annotations, the annotators discussed and resolved disagreements to reach a consensus for each article. This process revealed the need to modify the original taxonomy, as some operational risks, like process risk, were not well represented. To address this, the taxonomy was expanded by adding categories such as Process risk under Governance.

To reduce subjectivity, the annotators focused only on risks explicitly stated or clearly implied in the text that directly impacted the companies, excluding hypothetical or speculative risks. Each annotation included the risk class, family, specific type, a supporting quote (if available), and a brief justification of its relevance. After resolving disagreements, the final annotated dataset reflected the distribution of risks across the selected news articles as presented in Table 3. Twenty-one news items did not contain any identifiable risks.

At the Family level, the most frequently assigned categories were Government Business Policy (19.02%), Strategic Performance (11.68%), and Process Risk (8.15%), followed by Economic Outlook (7.61%) and Company Outlook (6.79%). Other notable families included Economic Variables (6.25%), Business Environment (4.08%), Brand

[1] https://www.corpusdata.org/intro.asp

Table 3 Distribution of risk classes in the annotated dataset

Risk class	Percentage (%)
Financial	31.79
Governance	29.35
Geopolitical	27.99
Social	7.07
Environmental	2.45
Technology	1.36

Perception and Competition (each 3.53%), and Litigation (3.26%), with lower-frequency categories such as Market Crisis, Trading Environment (2.99%), Human Capital (2.72%), and others like Political Violence, Climate Change, and Cyber.

Model Building and Results

As illustrated in Fig. 2, the inference pipeline comprised preprocessing, text embedding, retrieval of semantically similar annotated examples, and prompt construction, followed by model inference and post-processing. To evaluate the capability of LLMs for automated risk annotation from business news, we tested two state-of-the-art models: the open-source LLaMA−3.3–70B-Instruct and OpenAI's proprietary GPT-4o.

This comparison enables an assessment of performance differences between an open-access model and a commercial model in a complex multi-label classification task requiring fine-grained semantic interpretation of unstructured text. Both models were applied under identical experimental conditions using the same prompt structure and evaluation framework, allowing a controlled comparison of their effectiveness in identifying company-specific risks within the Cambridge Risk Taxonomy.

LLaMA was deployed on Finland's Puhti supercomputer. LLaMA can be deployed in a local or self-hosted environment, whereas GPT-4o is accessible only through OpenAI's cloud API. Both models analyzed business news using a unified prompt to identify company-specific risks and classify them under the Cambridge taxonomy. Fine-tuning was avoided due to the dataset's small size and imbalance across risk categories, making few-shot prompting a more practical approach.

Due to space constraints and the inclusion of lengthy full-text news articles, the complete prompts used for both models are only summarized in this paper. To ensure transparency and full reproducibility, the annotated dataset and detailed prompt examples are publicly available in the TurkuNLP GitHub repository.[2] The repository includes the full-text articles together with the annotations produced in this study. The summary of the full prompt used for both models (LLaMA−3.3–70B-Instruct and OpenAI's proprietary GPT-4o) was as follows:

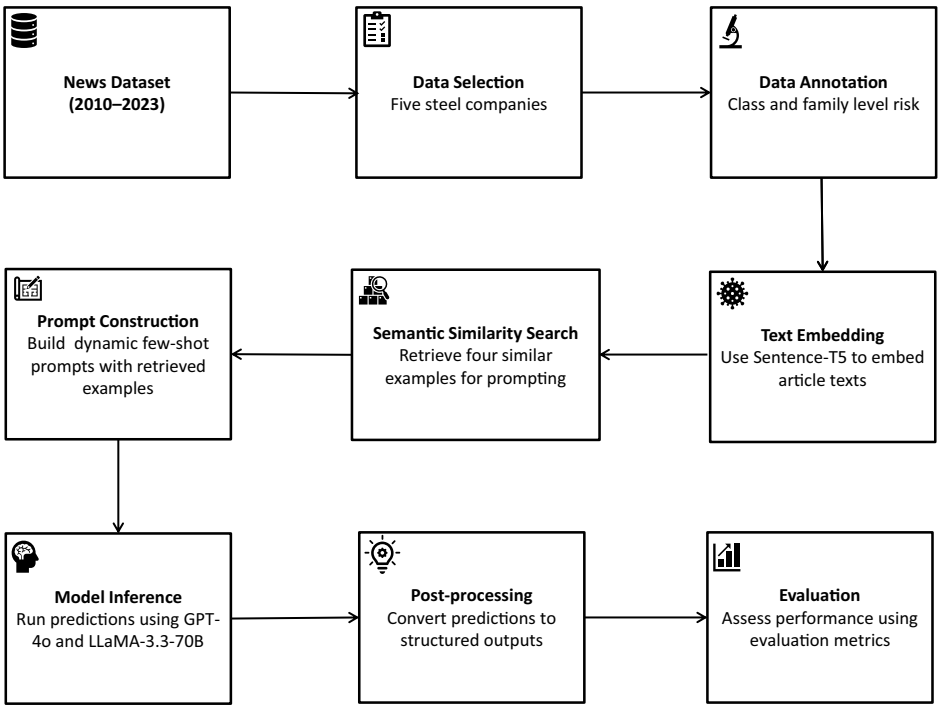

Fig. 2 Risk assessment pipeline

[2] https://github.com/TurkuNLP/AI-SIM.git

```
You are an insightful risk analyzer. You will be given a news
     article about a company named {Company}.

Your task is to:
1. Review the news and identify any risks that specifically
    affect {Company}'s business or operations.
2. Only consider risks that directly or indirectly impact {
    Company}. Do not include general risks mentioned in the
    news unless they have a clear and specific consequence
    for {Company}.
3. For each risk found:
    - Assign the correct Class and Family.
    - Provide a comment summarizing the risk.
    - Extract an exact quote from the news article that
       mentions or implies the risk (put the quote inside
       quotation marks "...").
    - Add a brief justification explaining why this is a risk
       .

Risk Classes and Families (with definitions):
[...followed by full taxonomy...]

Output Format:
Class: <Class>, Family: <Family>, Comment: <Your brief risk
    comment>, "<Exact quote from the news>", <Short
    justification>

### Examples:
{examples}

### New News:
News: {news_text}

### Annotation:
```

Semantic similarity refers to the process of measuring how closely related two terms are in meaning, even if they differ lexically. This is typically achieved by mapping the terms to a structured ontology and analyzing the relationships between them within that framework [29]. In our study, semantic similarity search was used to identify the most relevant examples for few-shot prompting. To enable this, we encoded each news article into a numerical vector using the Sentence-T5 model[3] [76], which captures the semantic content of the text. This embedding step was critical for selecting the most semantically similar annotated examples to include in each prompt. Subsequently, we applied a leave-one-company-and-year-out cross-validation strategy to ensure robust generalization and avoid data leakage.

This ensured that the model was never exposed to the same company or year during training and testing, thus avoiding data leakage and promoting generalization. Subsequently, this approach resulted in 48 folds. In each of the 48 resulting folds, articles from one company and one year were reserved as the test set, while the remainder formed the training set.

The prompt included the full text of the test news article and instructed the model to identify risks affecting the company's business or operations, assign the appropriate Class and Family according to the Cambridge Risk Taxonomy, and justify each annotation with supporting quotes and brief explanations. Few-shot prompting refers to including a small number of annotated examples in the prompt to guide the model's predictions. The few-shot examples were dynamically retrieved using Sentence-T5 embeddings to identify semantically similar annotated news articles.

Table 4 Overall performance metrics of the models across all folds

Metric (%)	GPT-4o	LLaMA 3.3
Average class-level F1-score (micro)	64%	58%
Average exact match accuracy	25%	9%
Average Hamming accuracy	64%	57%
Average Jaccard score	50%	46%
Average coverage	78%	78%

[3] https://huggingface.co/sentence-transformers/sentence-t5-base

To ensure realistic evaluation and avoid overly optimistic performance estimates, we applied a leave-one-company-and-year-out strategy, whereby examples originating from the same company and year as the test article were excluded from the prompt. This design encourages the model to generalize across companies and time periods, reflecting real-world deployment scenarios in which future news articles must be interpreted without access to near-duplicate examples.

The evaluation focused on the risk class level to ensure consistency and clarity, given the limited label variety. Model predictions were compared to human annotations using the micro-averaged F1 score for partial overlaps, exact match accuracy for full label agreement, and coverage to assess how often at least one correct risk class was predicted per article, acknowledging that exact matches are rare in multilabel tasks.

As presented in Table 4, the OpenAI model outperformed LLaMA 3.3 in risk analysis. As both models were implemented under identical conditions using the same prompts and evaluation setup, the superior performance of GPT-4o can be reasonably attributed to its larger parameter size and enhanced generalization capabilities compared to LLaMA 3.3. For the best model, the results across all folds showed an average micro F1 score of 64%, indicating that the model often identified the correct risk classes even when its predictions included additional classes.

The exact match accuracy was 25%, reflecting the strictness of this metric in a multilabel context where the model frequently predicted more classes than the human annotators, leading to mismatches even when many correct classes were included. The average coverage was 78%, demonstrating that in the majority of cases, the model successfully identified at least one relevant risk class for each article. Finally, the average Jaccard score was 50%, demonstrating that roughly half of the predicted labels overlapped with the human-annotated ones on a per-article basis. Together, these metrics reveal that while the model performs reliably in identifying many risks, it occasionally fails to capture the full set of relevant classes, particularly under strict evaluation criteria. These findings suggest that while the model tended to overpredict risks compared to human annotators, it was effective at recognizing relevant risk factors in a challenging, real-world dataset without requiring any fine-tuning.

Best Model Error Analysis and Output Reliability Assessment

This section evaluates the reliability of the best-performing model's outputs by examining its errors at the label level and, crucially, by assessing whether the model preserves the underlying risk narrative conveyed in the news articles.

Figure 3 presents a comparison between the annotated and predicted distributions of risk classes (Generated by the best model: GPT-4o) for ThyssenKrupp across the period 2016–2023. Each bar represents the number of risks within a given class identified in company-related news during a particular year. The left panel corresponds to the human-annotated data, while the right panel displays the model's predictions.

Overall, the model successfully captures the temporal evolution and categorical composition of risks reported for ThyssenKrupp. Both panels reveal similar fluctuations in risk frequency over time and indicate that financial, governance, and geopolitical risks are the dominant categories

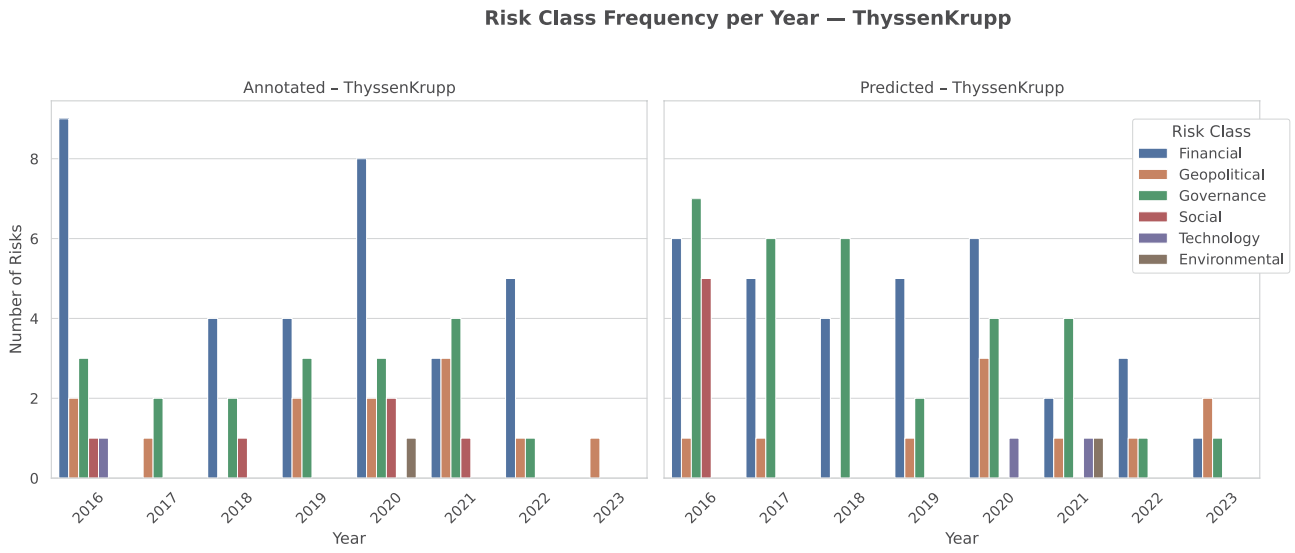

Fig. 3 Annotated and predicted risk class frequencies for ThyssenKrupp (2016–2023)

Table 5 Most frequently missed risk classes by the OpenAI model

Risk class	Times missed	Total occurrences
Geopolitical	20	73
Social	9	25
Governance	6	72
Financial	3	61
Environmental	2	9

affecting the company. The annotated data emphasises the predominance of financial risks, particularly during 2016, 2019, and 2020, whereas the predicted data shows a more balanced distribution between financial and governance risks and includes slightly higher frequencies of social and technology-related risks. This difference suggests that the model tends to interpret some textual cues more broadly, resulting in a modest overestimation of certain categories.

From a temporal perspective, both datasets indicate heightened risk activity between 2016 and 2020, a period marked by ThyssenKrupp's corporate restructuring, governance challenges, and exposure to market and trade fluctuations. A noticeable decline in risk frequency after 2021 reflects a reduction in media coverage or a stabilisation of company operations. Taken together, the comparison demonstrates that the model effectively reproduces the overall trajectory and relative prominence of risk classes over time, even though it exhibits some divergence in fine-grained class attribution. This pattern suggests that the model performs well in recognising the broader risk dynamics while displaying a tendency toward semantically inclusive classification.

Moreover, Table 5 shows that the OpenAI model most frequently missed Geopolitical risks, failing to detect them in 20 out of 73 cases. Many of these involved subtle or indirect references, such as regulatory shifts or cross-border policy issues, often embedded in economic or environmental contexts without clear geopolitical language. This implicitness, combined with overlap in linguistic features with Governance-related risks, likely contributed to confusion. The model's reliance on lexical cues makes it less effective at recognizing such nuanced risks, suggesting the need for clearer training examples and improved contextual prompts to enhance Geopolitical risk detection.

Table 6 presents three illustrative examples comparing the model's predicted risk classes with the true human-annotated classes. The original news articles were lengthy and information-dense; therefore, each item has been manually summarized to enhance readability and clarity of presentation. As shown, the LLM was generally effective in capturing core risks. However, it occasionally introduced additional risk classes not present in the ground truth.

We further evaluated the model at the level of narrative risk communication rather than solely at the level of individual labels. For each news article, both the human-annotated dataset and the model-predicted dataset include short free-text comments explaining why a given risk label applies to the focal company. To assess whether the model preserves the overall risk narrative of an article, we used GPT-4o-mini to generate concise, manager-oriented summaries based on the comments in each dataset. The paired summaries were produced using the same structured prompt, which instructed GPT-4o-mini to group risks by class and conclude with an overall assessment. Consequently, similarities in surface structure reflect the standardized summarization format rather than manual stylistic alignment.

Concretely, for every article, we provided the model with the list of risk classes, families, risk types, and the associated comments, together with an instruction to group the information by risk class and produce an executive-style summary of the main risks and their causes for the company. This procedure was applied twice per article, once using the human-annotated comments and once using the comments produced from the model's predictions, yielding a pair of summaries for each news item. Illustrative examples of these paired summaries are shown in Table 7. To evaluate the reliability of these summaries, we conducted a two-part analysis. First, we measured the lexical and semantic similarity between the human-aligned and model-generated summaries. Second, we assessed the quality of each summary independently.

We first examined the extent to which the two summaries for each article align lexically and semantically. Lexical

Table 6 Examples of model predictions vs. true annotations

News summary	True classes	Predicted classes
Odisha's decision to renew 18 expired mining leases, including those affecting Tata Steel, reflects governance and geopolitical risks linked to regulatory delays and legal actions that previously disrupted raw material access	Governance; Geopolitical	Governance; Financial
POSCO faced challenges as public protests erupted over the government's decision to divert Mahanadi river water for its proposed steel plant in Odisha, raising concerns about resource allocation and political opposition	Governance; Geopolitical	Geopolitical; Social; Environmental
ThyssenKrupp faced difficulties after organized cyberattacks originating from Southeast Asia compromised trade secrets and engineering data, threatening intellectual property and exposing the company to costly security breaches across its global operations	Technology; Financial	Technology

Table 7 Example comparison between predicted and annotated risk summaries

Predicted summary	Annotated (true) summary
Environmental: Tata faces risks related to environmental degradation, specifically concerning potential harm to endangered species and ecosystems. Increased industrial activity could lead to scrutiny from environmental groups, resulting in regulatory challenges and possible project delays as measures are taken to protect vulnerable wildlife.	*Environmental*: Tata faces risks related to climate change, specifically concerning the physical impacts on sensitive ecosystems, such as the Olive Ridley turtle nesting habitat. While the company received legal clearance, ongoing environmental pressures from NGOs could pose reputational and operational risks.
Geopolitical: The company is exposed to risks from government business policy, particularly regulatory scrutiny and potential changes in policy regarding environmental compliance. Disagreements among government bodies about compliance may lead to shifts in enforcement that could adversely affect project timelines and associated costs.	*Geopolitical*: The Company is exposed to risks from potential changes in government policies, particularly regarding environmental regulations. Activism from groups like Greenpeace highlights concerns over the selective enforcement of environmental laws that may favor Tata, suggesting that shifts in government stance could disrupt operations.
Governance: Allegations of regulatory non-compliance present significant governance risks for Tata. Reports of violations of the Forest Conservation Act could result in legal challenges, financial penalties, or operational disruptions, thereby impacting both the company's operations and its reputation.	*Governance*: Tata is currently involved in litigation related to trademark violations, which could lead to financial liabilities and distract from core business activities. Additionally, ongoing legal scrutiny concerning environmental practices could result in further regulatory challenges and impact the company's strategic direction.
Social: Tata is at risk from negative brand perception due to activism from organizations like Greenpeace, which campaigns against the company's projects. Such public opposition can damage the company's image and increase pressure from stakeholders to address the environmental issues raised.	*Social*: Tata's brand perception is at risk due to negative media coverage and public sentiment surrounding its environmental practices. Accusations of preferential treatment by the government could damage the company's reputation, leading to potential backlash from stakeholders and consumers.
Overall assessment: Tata's risk landscape is characterized by environmental, geopolitical, governance, and social challenges. The interplay of regulatory scrutiny, allegations of non-compliance, and public activism could significantly impact the company's operational efficiency and reputation. Proactive engagement and compliance strategies will be essential for mitigating these risks and maintaining stakeholder trust	**Overall assessment:** Tata is navigating a complex risk landscape characterized by environmental sensitivities, potential geopolitical shifts, and governance challenges. The interplay between regulatory scrutiny and public perception poses significant risks to the company's operations and reputation. Proactive management of these risks is essential to safeguard Tata's market position and ensure sustainable business practices

Table 8 Automatic evaluation metrics for summary quality

Metric	Type	Average Score
BERTScore (F1)	Semantic similarity	91%
ROUGE-1	Lexical overlap (unigrams)	55%
ROUGE-L	Lexical overlap (longest sequence)	30%
ROUGE-2	Lexical overlap (bigrams)	21%
BLEU	Lexical overlap (precision-based)	14%

similarity was evaluated using ROUGE and BLEU scores, while semantic similarity was assessed using BERTScore.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a widely used metric in machine translation and text summarization tasks. ROUGE-1 and ROUGE-2 quantify the lexical overlap of unigrams and bigrams between the generated and reference summaries, thereby capturing basic content similarity and local phrase coherence, while ROUGE-L measures the longest common subsequence of words, assessing the extent to which the generated summary preserves the sequential structure and syntactic fluency of the reference text [54].

BLEU (Bilingual Evaluation Understudy) is a language-independent metric that evaluates machine translation quality by measuring n-gram overlap between candidate and reference texts, incorporating a brevity penalty to approximate human judgments of adequacy and fluency [55].

In this study, ROUGE-1 and ROUGE-L yielded moderate scores, indicating partial lexical overlap through unigram and sequence-level matches, whereas ROUGE-2 and BLEU, both sensitive to n-gram precision, produced lower values, as expected for abstractive summarization tasks.

Given that lexical metrics are limited in their ability to capture meaning, semantic similarity was evaluated using BERTScore. BERTScore is an automatic metric designed to evaluate the quality of generated text by measuring its semantic similarity to a reference. Unlike traditional metrics that rely on exact word matches, BERTScore leverages contextual embeddings from pretrained language models (such as BERT) to capture the meaning of each token within its context. It calculates a similarity score between every token in the generated text and every token in the reference text, reflecting how closely their meanings align. Furthermore, BERTScore has been demonstrated to be more robust to paraphrasing and other linguistic variations, making it a more accurate and semantically informed measure of text quality [39].

BERTScore yielded a high semantic similarity between model-generated and human-derived summaries, with an average F1 of 91% across all articles as presented in Table 8.

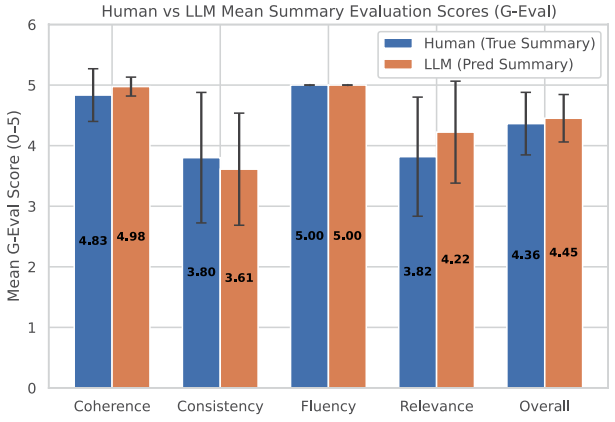

Fig. 4 Comparison of G-eval metrics between annotated and predicted summaries

This high score reflects strong semantic alignment between the model-generated summaries and those derived from the human annotations. The model consistently emphasises similar risk domains, such as Environmental, Geopolitical, Governance, and Social, and attributes them to comparable drivers, including regulatory scrutiny, NGO activism, and litigation. These results indicate that the system functions not only as a multi-label classifier but also as a generator of coherent, high-level risk briefings. Even when some fine-grained labels are misclassified, the resulting summaries remain closely aligned with the human-interpreted narratives. This supports the model's suitability for downstream applications that require an integrated view of a company's risk landscape rather than exact replication of every individual label. The full summary-generation prompt, paired summaries, and per-article BERTScore values are available in the TurkuNLP repository.[4]

As similarity metrics do not directly assess summary quality, we additionally employed G-Eval through the AdalFlow library.[5] G-Eval uses large language models with chain-of-thought reasoning and a structured evaluation template to assess coherence, consistency, fluency, and relevance. Prior work has shown that G-Eval, particularly when powered by GPT-4, aligns closely with human evaluations: for summarisation, it reaches a Spearman correlation of 0.514, outperforming previous automatic metrics [53].

In our study, G-Eval (implemented with GPT-4o as the evaluator) assessed two sets of summaries: those produced from human annotations ("True Summary") and those generated directly from model predictions ("Predicted Summary"). Each summary was evaluated against the risk statements from which it was constructed, allowing a systematic comparison of how well the model preserves semantic integrity and interpretive quality in both conditions.

The results, presented in Fig. 4, show that the LLM-generated summaries achieve near-parity with the annotation-based summaries across all four evaluation dimensions. Fluency and coherence received the highest scores (approximately 5.0 on a 0–5 scale), indicating that the summaries are structurally sound and grammatically well-formed. Slight differences in relevance and consistency suggest occasional generalisation or omission of minor details, which is expected in abstractive summarisation. Overall, the G-Eval results reinforce the conclusion that the model can produce coherent, accurate, and managerially useful risk narratives.

Taken together, these findings show that although the summaries differ lexically, they remain closely aligned in meaning and exhibit comparable qualitative characteristics. The model is therefore capable not only of predicting risk labels but also of producing coherent, managerially useful summaries that capture the essential elements of a company's risk exposure. This suggests that the approach holds value for practical applications where decision-makers require an integrated understanding of emerging risks rather than exact replication of each individual annotation.

Discussion

This study advances current research on automated risk detection by demonstrating how state-of-the-art LLMs can be used not only to identify risk labels in business news but also to generate coherent managerial summaries that reflect a company's overall risk environment. Compared with Pei et al. [34], who applied a flat taxonomy to short news excerpts, our work uses the hierarchical Cambridge Risk Taxonomy and long-form articles, enabling a more fine-grained understanding of complex risks such as Geopolitical and Governance-related issues. Moreover, rather than relying on fine-tuned transformer models, we evaluate more recent LLMs in a few-shot prompting setting, simulating realistic deployment scenarios where training data are limited.

In addition, recent work by Zhao et al. [64] has similarly examined the use of LLMs for supply chain risk identification from news, benchmarking several commercial models using Cambridge Risk Taxonomy. Their study demonstrates the feasibility of LLM-assisted pipelines for extracting and classifying risk events from real-world business articles. However, compared with their framework, our work makes several complementary contributions. First, we rely on a fully manually annotated dataset rather than LLM-assisted labeling, enabling more rigorous evaluation of model performance. Second, we extend the assessment beyond risk

[4] https://github.com/TurkuNLP/AI-SIM.git

[5] https://adalflow.sylph.ai/apis/eval/eval.g_eval.html

category classification by introducing a narrative reliability analysis, comparing LLM-generated managerial summaries with human-derived summaries. Third, our dataset is the first publicly available corpus of fine-grained, hierarchical risk annotations in an e-commerce–linked industrial context. Together, these differences position our results as a deeper investigation of how LLMs align with human reasoning, not only in categorical labels but also in higher-level managerial interpretations.

Therefore, several insights emerge from our experiments. First, LLMs can approximate human annotation at a general level, achieving moderate F1-scores and high coverage. However, they continue to struggle with domain-specific and context-dependent risks, especially Geopolitical risks, where cues are subtle or embedded in overlapping governance or economic narratives. This highlights the importance of domain knowledge and the need for clearer, more targeted examples in prompts.

Second, the models tended to overpredict risk classes, reflecting an inclination to generalize when contextual grounding is incomplete. This behavior emphasizes the sensitivity of few-shot prompting and the need for balanced example selection to avoid label inflation.

Third, our annotation process revealed substantial subjectivity in human risk interpretation. Even with a structured taxonomy, annotators disagreed on ambiguous cases, necessitating taxonomy refinements such as adding Process Risk. This underscores the inherent difficulty of abstract risk labeling and the importance of transparent guidelines for consistency.

A key methodological choice, constructing company-year-based folds, helped avoid data leakage and better emulate real-world forecasting scenarios. However, the length and complexity of news articles posed challenges for both human annotators and LLMs, suggesting that summarization or chunking methods should be explored in future work.

Importantly, we introduced an additional layer of evaluation that moves beyond label-level accuracy. By comparing manager-oriented summaries generated from human annotations and model predictions, we show that LLMs can reconstruct coherent risk narratives even when some fine-grained labels differ. The high semantic similarity indicates that the model preserves the core meaning of the human-derived risk story. Complementary G-Eval results further show that the model-generated summaries exhibit strong coherence, fluency, and relevance, closely mirroring the human-written summaries.

Similar to how Joglekar et al. [62] employed a "summary of summaries" approach to distill managerial insights from extensive textual information, and used G-Eval to demonstrate the quality and reliability of the generated summaries, our findings show that the LLM-generated summaries achieve a comparable condensation effect while maintaining strong evaluative quality. Although our approach does not aggregate multiple summaries into a meta-summary, the high semantic alignment between model- and human-generated narratives indicates that each summary already encapsulates condensed, decision-relevant information. This highlights the managerial relevance of LLM-generated risk summaries. While label-level accuracy remains imperfect, the strong semantic alignment between human and model-derived narratives demonstrates that LLMs can reliably convey the overarching themes and drivers of corporate risk exposure. For managers, such summaries offer a practical means of synthesizing complex news flows into structured risk intelligence that supports timely and informed decision-making. This dual capability analytical detection of risk categories and coherent narrative synthesis positions LLMs as promising components of scalable risk monitoring solutions in e-commerce and supply chain settings, where continuous awareness of geopolitical, governance, and societal developments is essential.

Finally, a striking illustration of why continuous risk monitoring matters is the failure of POSCO's $12 billion steel project in Odisha [77]. According to our annotated news, over more than a decade (2005–2017), POSCO encountered a confluence of regulatory, social, and operational risks, ranging from abrupt legal changes invalidating its expected mining rights to persistent local resistance against land acquisition. Despite significant investment and initial government support, the project was ultimately abandoned, resulting in reputational damage, ecological degradation, and lost economic opportunity. It also demonstrates how complex interactions between local communities, legal frameworks, and policy shifts can derail even the most strategically significant investments. Moreover, this case highlights the value of continuous risk assessment systems, especially in volatile and high-stakes industrial contexts. In other words, the risks were not invisible; they evolved gradually and were detectable to external observers. From this perspective, the POSCO case demonstrates why industries operating in politically sensitive, resource-intensive contexts require systematic, continuous, and fine-grained news-based monitoring.

Our study contributes to this need by showing that LLM-based annotation and summarisation can automatically extract and structure such signals from unstructured news text, capturing not only explicit risk mentions but also broader narrative patterns connected to the Cambridge Risk Taxonomy. While our model is not a complete early-warning system, the evidence indicates that it can serve as a foundational layer for proactive risk surveillance, highlighting emerging concerns, preserving company-level themes,

and maintaining coherent summaries aligned with expert interpretations.

Moreover, viewed through the lens of pervasive digital services, the proposed approach illustrates how LLM-based systems can support early risk awareness while enabling more inclusive access to decision-relevant insights. By generating coherent, manager-oriented summaries rather than raw analytical outputs, the system lowers cognitive and expertise barriers, allowing a broader range of organizational actors to engage with complex risk information.

Conclusion

This study evaluated the potential of LLMs to support automated risk identification and high-level interpretation of business news in e-commerce supply chains. Using a manually annotated dataset grounded in the hierarchical Cambridge Risk Taxonomy, we assessed GPT-4o and LLaMA 3.3–70B in a multi-label classification setting. Both models were able to detect a broad range of risk categories without fine-tuning, with GPT-4o achieving superior performance across F1-score, exact-match accuracy, and Jaccard similarity. These results illustrate the promise of LLMs for early-stage risk sensing while highlighting ongoing challenges related to domain specificity, prompt sensitivity, and inherent ambiguity in human risk annotation.

Beyond classification, we examined whether LLMs can generate coherent, decision-oriented summaries that approximate human interpretations of a company's risk situation. By comparing summaries derived from human annotations with those produced from model predictions, we showed that LLMs preserve the core semantic content of expert assessments. G-Eval results further confirmed strong coherence, fluency, and relevance, indicating that generated summaries are suitable for managerial contexts where understanding the broader risk landscape is more important than exact label reproduction. Taken together, our findings demonstrate that LLMs are promising tools for proactive and scalable risk monitoring. They can operate at two complementary levels: (1) analytical: identifying specific risk classes, and (2) managerial: summarizing the overall risk landscape in a coherent narrative. This dual capability positions LLM-based systems as valuable components of intelligent, risk-aware decision-support tools in complex, data-rich environments such as e-commerce and global supply chains.

To the best of our knowledge, no publicly available dataset offers similarly detailed, hierarchical risk annotations for long-form e-commerce and supply chain–related news. Developing such a corpus was therefore essential to enable systematic evaluation and reproducibility in this domain. While the absence of comparable datasets limits direct benchmarking, our work establishes a foundation for future research on domain-aware risk monitoring, annotation guidelines, prompt engineering for subjective tasks, and integration of summarization techniques for lengthy or complex texts.

In conclusion, we answered the three research questions by constructing a taxonomy-based risk dataset, empirically assessing LLMs' ability to classify risks from real-world news in e-commerce–enabled supply chains, and showing that their generated summaries retain the essential managerial insights conveyed by human annotations. These findings underscore the potential of LLM-based systems to enhance proactive risk sensing in volatile global supply chains. By combining structured risk labels with coherent narrative summaries, LLMs offer a scalable pathway for organizations to monitor evolving risks, anticipate disruptions, and support informed managerial decision-making. Moreover, our results demonstrate how LLM-enabled risk monitoring can be operationalized as a pervasive digital service, enabling more accessible, timely, and responsible decision support within e-commerce-enabled supply chains.

Future Research Directions

This study opens several promising directions for future research. Because the current analysis focuses on steel companies operating in e-commerce–related contexts, future work could examine additional industries with different risk profiles, enabling broader generalization across sectors. Expanding research to domains where Environmental, Social, or Operational risks are more prominent would deepen understanding of how LLMs behave in varied supply chain environments. The dataset used in this study, while sufficient for the exploratory aims, also points toward opportunities for developing larger, more balanced corpora that capture a wider range of risk events. Increasing the volume and diversity of annotated news articles would support more comprehensive evaluations of model performance and improve the detection of infrequent but strategically important risk types.

In addition, although the Cambridge Risk Taxonomy offers a robust foundation for high-level business risk analysis, future research could benefit from integrating supply chain–specific attributes and operational risk factors. A more granular taxonomy would enhance the precision with which LLMs can identify risks across complex supply networks. The study also highlights the interpretive nature of risk annotation. Future work could investigate methods for increasing annotation consistency, such as multi-annotator adjudication protocols, active learning support, or the use of

hybrid human–LLM annotation pipelines, to reduce subjectivity and improve reproducibility.

Furthermore, the present study relied on prompt-based inference without model fine-tuning, reflecting realistic early-stage deployment scenarios where domain-specific training data may be limited. Future research could explore fine-tuned or instruction-optimized models, particularly for long and information-dense news articles where domain adaptation may lead to measurable performance improvements.

Since the dataset consists of publicly available news articles published between 2010 and 2023, it is possible that some texts may overlap with the pretraining corpora of large language models. However, the human annotations based on the Cambridge Risk Taxonomy were newly created as part of this study and were not included in model training. The task requires semantic alignment with a structured risk ontology rather than simple memorization of textual content. Nevertheless, future work could evaluate the approach on proprietary or newly collected documents to further minimize potential overlap with pretraining data.

An important insight from this study is that LLMs may misclassify fine-grained labels while still generating coherent, semantically aligned risk narratives. This suggests a promising direction for shifting evaluation frameworks toward narrative-level or story-based assessments that better reflect managerial needs. Future work could design and validate such narrative-centric evaluation metrics to complement traditional label-level accuracy.

Looking ahead, to better tailor LLM-based risk detection to supply chain applications, future datasets should incorporate supply chain–relevant metadata, including: (1) the company's role in the supply network (supplier, manufacturer, distributor); (2) geographic location of suppliers and facilities; (3) known supplier–buyer relationships; (4) risk propagation pathways across the network; and (5) a refined taxonomy designed specifically for supply chain risk, capturing issues such as logistics delays, supplier insolvency, material shortages, and regulatory compliance failures. Collectively, these avenues point toward the creation of more context-aware, risk-sensitive AI systems capable of supporting strategic supply chain decision-making through both precise risk detection and managerially meaningful summaries.

Author Contributions Laleh Davoodi: Conceptualization, Methodology, Data analysis and model building, Data annotation and curation, Writing – Original Draft, Writing – Review and Editing. Filip Ginter: Conceptualization, Methodology, Data annotation guidance, Writing – Review and Editing, Supervision. Sima Salimi: Data annotation, Writing – Review and Editing. Harri Lorentz: Conceptualization, Data annotation guidance, Writing – Review and Editing, Supervision.

Funding Open Access funding provided by University of Turku (including Turku University Central Hospital). We gratefully acknowledge the financial support provided by Business Finland, as this research was carried out within the framework of the AI-SIM project. We also thank CSC – IT Center for Science for their support, particularly for providing the computing resources used to run the LLMs.

Data Availability The data supporting the findings of this study are available on the TurkuNLP GitHub repository. https://github.com/TurkuNLP/AI-SIM.git

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

Research Involving Human and/or Animals Not applicable.

Informed Consent Not applicable.

References

1. Kabus J, Miciuła I, Piersiala L. Risk in Supply Chain Management. Eur Res Stud J. XXII. 2020;I(4):467–80. https://doi.org/10.35808/ersj/1694.
2. Christopher M, Peck H, Abley J, Haywood M, Saw R, Rutherford C, Strather M. Creating resilient supply chains: a practical guide. Cranfield University, Bedford, UK, 2003. https://dspace.lib.cranfield.ac.uk/bitstream/1826/4374/1/Creating_resilient_supply_chains.pdf.
3. Zsidisin GA. A grounded definition of supply risk. J Purchas Supply Manage. 2003;9:217–24. https://doi.org/10.1016/j.pursup.2003.07.002.
4. Harland C, Brenchley R, Walker H. Risk in supply networks. J Purchas Supply Manage. 2003;9(2):51–62. https://doi.org/10.1016/S1478-4092(03)00004-9.
5. Mitchell VW. Organizational risk perception and reduction: a literature review. Br J Manag. 1995;6:115–33. https://doi.org/10.1111/j.1467-8551.1995.tb00089.x.
6. Zsidisin, G. A., Panelli, A., Upton, R. Purchasing organization involvement in risk assessments, contingency plans, and risk management: an exploratory study. Supply Chain Manage Int J. 2000;5(4):187–97. https://doi.org/10.1108/13598540010347307.
7. Caldwell N, Harland C, Powell P, Zheng J. Impact of e-business on perceived supply chain risks. J Small Bus Enterp Dev. 2013;20(4):688–715. https://doi.org/10.1108/JSBED-12-2011-0036.

8. Tang CS. Perspectives in supply chain risk management. Int J Prod Econ. 2006;103(2):451–88. https://doi.org/10.1016/j.ijpe.2005.12.006.
9. Mentzer JT, DeWitt W, Keebler JS, Min S, Nix NW, Smith CD, et al. Defining supply chain management. J Bus Logist. 2001;22(2):1–25. https://doi.org/10.1002/j.2158-1592.2001.tb00001.x.
10. Blos MF, Quaddus M, Wee HM, Watanabe K. Supply chain risk management (SCRM): a case study on the automotive and electronic industries in Brazil. Supply Chain Manage Int J. 2009;14(4):247–52. https://doi.org/10.1108/13598540910970072.
11. Clark TH, Lee HG. Performance, interdependence and coordination in business-to-business electronic commerce and supply chain management. Inf Technol Manage. 2000;1(1):85–105. https://doi.org/10.1023/A:1019108621684.
12. Abdinazar K. Growth in European e-commerce: analyzing the surge in online shopping and consumer behaviour. GSJ. 2025;13(2):2320–9186.
13. Chen J, Sohal AS, Prajogo DI. Supply chain operational risk mitigation: a collaborative approach. Int J Prod Res. 2013;51(7):2186–99. https://doi.org/10.1080/00207543.2012.727490.
14. Schmenner RW, Swink ML. On theory in operations management. J Oper Manag. 1998;17(1):97–113. https://doi.org/10.1016/S0272-6963(98)00028-X.
15. Gupta A. E-Commerce: role of E-Commerce in today's business. Int J Comput Corporate Res. 2014;4(1):1–8. ISSN (Online): 2249-054X.
16. Srivastava A. The application & impact of artificial intelligence (AI) on E-commerce. Contemp Issues Commerce Manage. 2021;1(1):165–75 (ISBN: 978-93-83569-10-6).
17. Urrea NT, Vishkaei BM, De Giovanni P. Operational risk management in e-commerce: a platform perspective. IEEE Trans Eng Manage. 2024;71:3807–19. https://doi.org/10.1109/TEM.2024.3358717.
18. Shahsavari M, Hussain OK, Saberi M, Sharma P. Empowering supply chains resilience: LLMs-powered BN for proactive supply chain risk identification. 2024.
19. Jin B, Sun Q, Chen L. Enhancing supply chain transparency in emerging economies using online contents and LLMs. In: 2025 international conference on information networking (ICOIN), 2025;487–492.
20. Shahsavari M, Hussain OK, Saberi M, Sharma P. Event identification for supply chain risk management through news analysis by using large language models. The Review of Socionetwork Strategies. 2024;18(2):255–78. https://doi.org/10.1007/s12626-024-00169-z.
21. Zhao M, Hussain O, Zhang Y, Saberi M, Leshob A. Enhancing supply chain risk management with large language models: software prototyping and interactive visualization. In: Proceedings of the 2024 IEEE international conference on e-business engineering (ICEBE), 2024;284–291. https://doi.org/10.1109/ICEBE62490.2024.00051.
22. Cambridge Centre for Risk Studies. Cambridge Taxonomy of Business Risks (Version 2.0). University of Cambridge Judge Business School. Retrieved from 2019. https://www.jbs.cam.ac.uk/faculty-research/centres/risk/publications/.
23. Zhu B, Vuppalapati C. Enhancing supply chain efficiency through retrieve-augmented generation approach in large language models. In: Proceedings of 2024 IEEE 10th international conference on big data computing service and machine learning applications (BigDataService), 2024;117–121. https://doi.org/10.1109/BigDataService62917.2024.00025.
24. Quan Y, Liu Z. InvAgent: A large language model-based multi-agent system for inventory management in supply chains (preprint). 2024. https://doi.org/10.48550/arXiv.2407.11384.
25. Teixeira AC, Marar V, Yazdanpanah H, Pezente A, Ghassemi M. Enhancing credit risk reports generation using LLMs: An integration of Bayesian networks and labeled guide prompting. In: Proceedings of the fourth ACM international conference on AI in finance, 2023;340–348. https://doi.org/10.1145/3604237.3626902.
26. Koli, L., Kalra, S., Thakur, R., Saifi, A., Singh, K. (2025). AI-Driven IRM: Transforming insider risk management with adaptive scoring and LLM-based threat detection (preprint). https://doi.org/10.48550/arXiv.2505.03796.
27. Sukumar A, Edgar D, Grant K. An investigation of e-business risks in UK SMEs. Manage Sustain Dev. 2011;7(4):380–401. https://doi.org/10.1504/WREMSD.2011.042892.
28. Liu X, Ahmad SF, Anser MK, Ke J, Irshad M, Ul-Haq J, et al. Cybersecurity threats: a never-ending challenge for e-commerce'. Front Psychol. 2022;13:927398. https://doi.org/10.3389/fpsyg.2022.927398.
29. Hliaoutakis A, Varelas G, Voutsakis E, Petrakis EG, Milios E. Information retrieval by semantic similarity. Int J Semantic Web Inf Syst. 2006;2(3):55–73. https://doi.org/10.4018/jswis.2006070104.
30. Statista. Total retail e-commerce revenue worldwide in 2025, by region (in billion U.S. dollars) [Graph]. In Statista. Retrieved July 21, 2025, from, 2025. https://www.statista.com/forecasts/1117851/worldwide-e-commerce-revenue-by-region.
31. Caldara D, Iacoviello M. Measuring geopolitical risk. American Economic Review. 2022;112(4):1194–225. https://doi.org/10.1257/aer.20191823.
32. Jawadi F, Rozin P, Gnegne Y, Cheffou AI. Geopolitical risks and business fluctuations in Europe: a sectorial analysis. Eur J Political Econ. 2024;85:102585. https://doi.org/10.1016/j.ejpoleco.2024.102585.
33. Gaurav M. "Building Supply Chain Resilience Using Artificial Intelligence in Risk Management Systems. Smart Services Summit 2023;55–67. https://doi.org/10.1007/978-3-031-60313-6_5.
34. Pei J, Vadlamannati S, Huang LK, Preoţiuc-Pietro D, Hua X. Modeling and detecting company risks from news. In: Proceedings of NAACL 2024: human language technologies (industry track), 2024;63–72.
35. Simchi-Levi D, Kaminsky P, Simchi-Levi E. Managing the supply chain: the definitive guide for the business professional. McGraw-Hill. ISBN: 0071435875,0071410317. 2004.
36. Dutta P, Suryawanshi P, Gujarathi P, Dutta A. Managing risk for e-commerce supply chains: an empirical study. IFAC-PapersOnLine. 2019;52(13):349–54. https://doi.org/10.1016/j.ifacol.2019.11.143.
37. Hillman M, Keltz H. Managing risk in the supply chain: a quantitative study. AMR Res. 2007;1–24.
38. Laudon KC, Traver CG. E-Commerce Business Technology Society (Thirteenth). ISBN: 9781282693159. 2017.
39. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating text generation with BERT (arXiv preprint). 2019. https://doi.org/10.48550/arXiv.1904.09675.
40. Venkatesh VG, Rathi S, Patwa S. Analysis on supply chain risks in Indian apparel retail chains and proposal of risk prioritization model using Interpretive structural modeling. J Retail Consum Serv. 2015;26:153–67. https://doi.org/10.1016/j.jretconser.2015.06.001.
41. Üstündağ A, Çıkmak S, Eyiol MÇ, Ungan MC. Evaluation of supply chain risks by fuzzy DEMATEL method: a case study of iron and steel industry in Turkey. International Journal of Production Management and Engineering. 2022;10(2):195–209. https://doi.org/10.4995/ijpme.2022.17169.
42. Xiong G, Helo P. Challenges to the supply chain in the steel industry. Int J Logistics Econ Global. 2008;1(2):160–75. https://doi.org/10.1504/IJLEG.2008.020529.

43. Sandhu M, Helo P, Kristianto Y. Steel supply chain management by simulation modelling. Benchmarking Int J. 2013;20(1):72–91. https://doi.org/10.1108/14635771311299 4.
44. World Steel Association. The role of steel manufacturing in the global economy. Retrieved from 2019. https://worldsteel.org/wp-content/uploads/The-role-of-steel-manufacturing-in-the-global-economy.pdf
45. Xu J, Yu Q, Hou X. Sustainability assessment of steel industry in the belt and road area based on DPSIR model. Sustainability. 2023;15(14):11320. https://doi.org/10.3390/su151411320.
46. Harandi Amin M, Sharifi-Ghazvini M, Aghasi S. Designing a resilient supply chain model with a focus on financial outcomes (case study: mobarakeh steel company). Digital Transform Administ Innov. 2024;2(2):79–88. https://doi.org/10.61838/dtai.2.2.9
47. Kaplas O. Sustainability challenges in steel supply chain: Identifying and addressing key factors in project business context (Master's thesis, Tampere University Repository), 2024. https://urn.fi/URN:NBN:fi:tuni-202411019791.
48. Ramadhan FA, Mansur A, Setiawan N, Salleh MR. "An analytical risk mitigation framework for steel fabrication supply chains using fuzzy inference and house of risk" in Supply Chain Analytics, 2025;10. https://doi.org/10.1016/j.sca.2025.100122.
49. Liao J, Deng S, Huan X, Cooper D. Bayesian model selection for network discrimination and risk-informed decision-making in material flow analysis. J Indus Ecol. 2025. https://doi.org/10.1111/jiec.70034.
50. Choi JW, Shin KS. A Study on the risk identification and management of global supply chain risks: focusing on the domestic steel industries. 2025;35(1):71–84. https://doi.org/10.17825/klr.2025.35.1.71
51. Singhal P, Agarwal G, Mittal ML. Supply chain risk management: Review, classification and future research directions. Int J Business Sci Appl Manage. 2011;6(3).
52. Ivanov D, Dolgui A, Sokolov B. The impact of digital technology and Industry 4.0 on the ripple effect and supply chain risk analytics. Int J Prod Res. 2019;57(3):829–46. https://doi.org/10.1080/00207543.2018.1488086.
53. Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C. "G-Eval: NLG evaluation using GPT-4 with better human alignment" (arXiv preprint). 2023. https://doi.org/10.48550/arXiv.2303.16634.
54. Lin CY. ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out (workshop), 2004;74–81.
55. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics, 2002;311–318.
56. Dale R. GPT-3: What's it good for? Nat Lang Eng. 2021;27(1):113–8. https://doi.org/10.1017/S1351324920000601.
57. Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, et al. A comprehensive overview of large language models. ACM Trans Intell Syst Technol. 2025;16(5):1–72. https://doi.org/10.1145/3744746.
58. Yenduri G, Srivastava G, Maddikunta PKR, Jhaveri RH, Wang W, Vasilakos AV, Gadekallu TR. Generative pre-trained transformer: a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions (preprint). 2023. https://doi.org/10.48550/arXiv.2305.10435.
59. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave É, Lample G. LLaMA: Open and efficient foundation language models (arXiv). 2023. https://doi.org/10.48550/arXiv.2302.13971.
60. Bharathi MG, Prasanna Kumar R, Vishal Krishh P, Keerthinathan A, Lavanya G, Meghana MKU, et al. An analysis of large language models: their impact and potential applications. Knowl Inf Syst. 2024;66(9):5047–70. https://doi.org/10.1007/s10115-024-02120-8.
61. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure (preprint). 2022. https://doi.org/10.48550/arXiv.2203.05794
62. Joglekar N, Nie W, Alem Fonseca M, Tsolakis N, Kumar M. An AI-driven approach to assess sentiments and interpret context in a critical mineral supply chain. Int J Prod Res. 2025;1–29. https://doi.org/10.17863/CAM.119174.
63. Jannelli V, Schoepf S, Bickel M, Netland T, Brintrup A. Agentic LLMs in the supply chain: towards autonomous multi-agent consensus-seeking (arXiv preprint). 2024. https://doi.org/10.48550/arXiv.2411.10184.
64. Zhao M, Hussain O, Zhang Y, Saberi M, Leshob A. "Benchmarking large language models for supply chain risk identification: an extended evaluation within the LARD-SC framework. In: Service oriented computing and applications, 2025;1–19. https://doi.org/10.1007/s11761-025-00474-7.
65. Wang Z, Pang Y, Lin Y, Zhu X. Adaptable and reliable text classification using large language models. In: 2024 IEEE International conference on data mining workshops (ICDMW), 2024;7–74. https://doi.org/10.1109/ICDMW65004.2024.00015.
66. Rouzegar H. LLM-powered active learning for cost-effective text classification (Doctoral dissertation). 2024. https://hdl.handle.net/10155/1867.
67. Kostina A, Dikaiakos MD, Stefanidis D, Pallis G. Large language models for text classification: Case study and comprehensive review (arXiv preprint). 2025. https://doi.org/10.48550/arXiv.2501.08457.
68. Lin CY, Tseng TL, Xu H. GPT-augmented Bayesian reinforcement learning framework for multi-objective supplier selection. IEEE Trans Eng Manage. 2025. https://doi.org/10.1109/TEM.2025.3603183.
69. Quan Y, Xu Y, Chen G, Benaben F, Montreuil B."Leveraging large language models for risk assessment in hyperconnected logistic hub network deployment (preprint). 2025. https://doi.org/10.48550/arXiv.2503.21115.
70. Ji Z, Yu T, Xu Y, Lee N, Ishii E, Fung P. Towards mitigating LLM hallucination via self reflection. In: Findings of ACL: EMNLP 2023, 2023;1827–1843.https://doi.org/10.18653/v1/2023.findings-emnlp.123.
71. Arabzadeh N, Clarke CL. A human-AI comparative analysis of prompt sensitivity in LLM-based relevance judgment. In: Proceedings of SIGIR 2025, 2025. https://doi.org/10.1145/3726302.3730159.
72. Peters U, Chin-Yee B. Generalization bias in large language model summarization of scientific research. R Soc Open Sci. 2025;12(4):241776. https://doi.org/10.1098/rsos.241776.
73. Stefanus A, Hartono M. A framework for evaluating the performance of supply chain risk in e-commerce. In: Proceedings of IEOM 2018 Bandung, Indonesia, pp. 1887–1894. 2018. http://repository.ubaya.ac.id/id/eprint/34245.
74. Gottlieb S, Ivanov D, Das A. Case studies of the digital technology impacts on supply chain disruption risk management. Logistik im Wandel der Zeit - Festschrift für Wolfgang Kersten. 2019;23–52. https://doi.org/10.1007/978-3-658-25412-4_2.
75. European Steel Association (EUROFER). European Steel in Figures 2025. 2025. https://www.eurofer.eu/assets/publications/brochures-booklets-and-factsheets/european-steel-in-figures-2025/European-Steel-in-Figures-2025_23062025.pdf.

76. Ni J, Abrego GH, Constant N, Ma J, Hall K, Cer D, Yang Y. Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models. In: Findings of ACL 2022, 2022;1864–1874. https://doi.org/10.18653/v1/2022.findings-acl.146.
77. Satapathy D. Posco closes last chapter in Odisha project. Business Standard. 2017. https://www.business-standard.com/article/companies/posco-closes-last-chapter-in-odisha-project-117031700810_1.html.