



An AI-driven approach to assess sentiments and interpret context in a critical mineral supply chain

Nitin Joglekar, Wei Nie, Mariel Alem Fonseca, Naoum Tsolakis & Mukesh Kumar

To cite this article: Nitin Joglekar, Wei Nie, Mariel Alem Fonseca, Naoum Tsolakis & Mukesh Kumar (22 Jul 2025): An AI-driven approach to assess sentiments and interpret context in a critical mineral supply chain, International Journal of Production Research, DOI: [10.1080/00207543.2025.2532142](https://doi.org/10.1080/00207543.2025.2532142)

To link to this article: <https://doi.org/10.1080/00207543.2025.2532142>



© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



[View supplementary material](#)



Published online: 22 Jul 2025.



[Submit your article to this journal](#)



Article views: 1118



[View related articles](#)



[View Crossmark data](#)



Citing articles: 1 [View citing articles](#)



An AI-driven approach to assess sentiments and interpret context in a critical mineral supply chain

Nitin Joglekar ^a, Wei Nie ^b, Mariel Alem Fonseca ^b, Naoum Tsolakis ^b and Mukesh Kumar ^b

^aVillanova School of Business, Villanova University, Villanova, PA, United States; ^bIndustrial Resilience Research Group, Department of Engineering, University of Cambridge, Cambridge, United Kingdom

ABSTRACT

This exploratory article argues against using Large Language Models (LLMs) as a ‘black box’, without human scrutiny, for generating interpretable context while assessing sentiment metrics. These metrics support supply chain (SC) decision-making under information asymmetry and public policies in critical mineral networks. Using a dataset of 5,168 news articles before and after a truck strike in the Democratic Republic of Congo (DRC), we enumerate observed sentiment shifts across cobalt SC echelons (i.e., DRC, China, U.S.A. and Canada) using an LLM with long context windows. As shifts can be small, assessments need to be precise, significant, and interpretable. To this end, we devise and deploy a ‘Context Enhanced Supply Chain Sentiment and Summaries’ (CESCSS) framework to provide reliable and interpretable outcomes. Through statistical testing, our findings indicate that context matters in assessing sentiment shifts across SC operations echelons. Findings also illustrate differences in sentiments and offer context summary-based interpretations for these differences based on end-to-end information asymmetries. In addition, results showcase the reliability of human ratings and then demonstrate that human assessments are statistically equivalent to context-enhanced LLM sentiment valuations. We discuss pathways for applying the CESCSS framework toward theory development and managerial decisions in critical mineral SCs.

ARTICLE HISTORY

Received 13 November 2024
Accepted 28 May 2025

KEYWORDS

Long context LLMs; critical mineral supply chains; context and sentiment; reliability and interpretability; context enhanced supply chain sentiment and summaries

1. Introduction

The advent of the World Wide Web provided the opportunity to publish news articles that are immediately accessible worldwide, serving as a valuable and abundant data source. Indicatively, the average number of news articles published daily in different countries (in English) can reach as high as 10,000 (Hamborg, Meuschke, and Gipp 2020), resulting in unprecedented data availability that can be leveraged to inform decision-making. Indicatively, the supply chain (SC) management domain has gradually evolved to recognise news articles’ potential roles and applications within research and practical contexts (Chae 2015). In SC management, analytics form the foundation for collecting, storing, and analysing diverse evidence to support data-driven decision-making for optimising performance and ensuring continuous improvement across end-to-end operations (Tiwari et al. 2024). For example, historical data are used for data-driven decision-making in dynamic pricing and inventory control problems under complete/incomplete demand information (Chen et al. 2021).

Nevertheless, despite the availability of news sources, the lack of information sharing among multi-echelon SC stakeholders reduces efficiency owing to information deficits, impacting trust (Inderfurth, Sadrieh, and Voigt 2013). Additionally, despite an increasing body of studies concerning the utilisation of data and analytical capabilities pertinent to SC management (Hazen et al. 2014; Trkman et al. 2010), the predominant emphasis has remained on conventional data sources (such as enterprise resource planning) and analytical techniques (including optimisation algorithms and forecasting) that influence SC planning and execution. Textual data derived from news sources have only recently been utilised in a SC context, e.g., to predict demand volatility in pharmaceuticals during periods of crisis (Nguyen et al. 2022).

Notwithstanding the technical advancements in extracting context from textual data, most of the respective outputs and algorithm-induced decisions are not straightforward due to the opaqueness of the techniques employed (Kosasih and Brintrup 2022). To this end,

CONTACT Mukesh Kumar  mk501@cam.ac.uk  Industrial Resilience Research Group, Department of Engineering, University of Cambridge, 17 Charles Babbage Rd, Cambridge CB3 0FS, United Kingdom

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/00207543.2025.2532142>.

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

decisions relying on raw textual data may be based on incomplete or misunderstood information without interpretation, leading to less effective outcomes. Recently, advancements in Generative Artificial Intelligence (GenAI), mainly through Large Language Models (LLMs), have amplified interest in the applicability of these digital technologies in operations management. LLM-based technology systems are used to explain SC decisions (e.g., by identifying productivity outliers in a network of multiple nodes, monitoring shipping cost spikes, or tracking shifts in demand), further providing the capability to support end-to-end decision-making scenarios and significantly reducing the time to make decisions (Menache et al. 2025). For instance, Jackson, Jesus Saenz, and Ivanov (2023) provided guidelines on the human-AI collaboration for the automatic development of simulation models based on verbal descriptions. To a further extent, Jackson et al. (2024) proposed a framework that pinpoints the areas and methods for applying GenAI within SC management, emphasising improvements in decision-making, process optimisation, investment prioritisation, and skills development.

Decisions related to SCs, as reported in various news sources, are being analysed by LLMs to generate sentiment assessments. Sentiments are defined as metrics with a negative, neutral, or positive polarity around an observed corpus's 'prevailing attitudes' (Sonnier, McAlister, and Rutz 2011). For instance, event studies investigated the tweet sentiment before, during, and after SC glitches and showed that such posts influence stock market returns, illustrating the potential of social media to amplify the visibility of SC issues (Schmidt et al. 2020). LLMs' training data in this regard not only accounts for widely recognised dictionaries, e.g., developed by Loughran and McDonald (2011) to evaluate sentiment, but also brings in context while assigning weights based on attention to critical elements of news items (Vaswani et al. 2017).

LLMs have been drawn upon to assess sentiment in challenging situations such as geopolitical events and capital market shifts. Notably, the US President's communication on Twitter significantly increased the predictive power of oil price prediction models (Beyer Díaz et al. 2024). To this end, extracting the content of textual data to identify sentiments, emotions, and connotations provides informative signals and insights for SC management in terms of assessing responses to risks (Wu 2023) and predicting demand (Anderson, Mahar, and Modi 2024). Nonetheless, LLMs may exhibit biases or hallucinations (Gupta et al. 2024; Venkit et al. 2024), i.e., "... the phenomenon where the model generates false information not supported by the input" (Xiao and Wang 2021). Especially in SC management, ensuring the LLM-generated

output's interpretability, i.e., the ability to explain quantitative outcomes to provide context for decision-makers, has been acknowledged as a critical issue (Baryannis et al. 2019). The quality of the analytical output produced by such AI-based systems can facilitate data-driven, objective decision-making for achieving SC resilience (Belhadi et al. 2021; Dey et al. 2023). Hence, detecting sentiment shifts and generating interpretability is especially important for the resilience of critical resource SCs owing to disruptive events' complexity and dynamic nature (Kleindorfer and Saad 2005), with ramifications that can propagate across end-to-end operations.

Therefore, the aim of this research is to propose a reliable and interpretable approach for exploring the intricate relationship between context and sentiment shifts in textual data that can inform decision-making for SC resilience. To this effect, this research attempts to address the following research questions (RQs) with underlying theory issues being explicated in Section 2:

- RQ#1 – *What reliable LLM-based method can be used to extract context from textual data and enhance the interpretability of SC sentiment analysis?*
- RQ#2 – *How can sentiment shifts estimated by an LLM be used to detect disruptions' propagation across SC echelons?*
- RQ#3 – *Can LLM-based sentiment analysis reflect shifts based on information asymmetry in SCs?*

Our work argues against using LLMs as a 'black box' without context and human scrutiny for assessing sentiment metrics. We recognise the possibility of hallucinations in LLM outputs and the need for a mechanism to examine these outcomes before reporting them. Indeed, contemporary research underscores that LLMs can generate compelling, albeit potentially misleading, explanations as they typically use strategies such as selective evidence or question reframing without transparent, interpretable reasoning processes (Ajwani et al. 2024). Hence, in response to the enunciated RQs, we developed a mixed-methods research approach, resulting in the proposed framework termed 'Context Enhanced Supply Chain Sentiment and Summaries' (CESCSS), which facilitates more robust findings and inspires theory development when examining complex dynamic phenomena (Golicic and Davis 2012). Our research approach is informed by the Syntactic and Semantic Attention Summary (SSAS) methodology (Mayande et al. 2024). The SSAS methodology argues for creating summaries to capture a dataset's diverse context, which is then used as input prompts to generate sentiment.

In particular, to address our research queries, the object of scrutiny was a set of 5,168 news articles (with

43 million words in total, in English) concerning different timeframes of a disruptive event to a critical mineral SC. Especially, this research leveraged the Factiva news database and Google's Gemini 1.5 Flash, an LLM with long context windows (i.e., one million input tokens), to perform context-driven sentiment analysis. We used automated and manual sentiment reliability and interpretation techniques to observe statistically significant differences. In contrast to traditional and static dictionary-based methods like Loughran and McDonald (2011), we used Google's Gemini LLM, which dynamically interprets sentiment based on its own training data that generated dictionaries, together with its advanced understanding of context. This strategy allowed for nuanced sentiment extraction, making it especially effective for analysing disruptions in critical mineral SCs. To measure the sentiment, we used a scale of -1 to 1 employed in other studies (Bau, Hermila, and Latief 2024). Thereafter, we deployed a detailed Design of Experiments to analyse shifts in sentiments before and after the studied disruptive event and across different geographically located operations echelons.

Overall, this research contributes to the Operations Management research community by taking initial steps toward developing a reliable approach for generating context-driven sentiments in SCs for decision-making. The proposed CESCSS framework can set up pathways toward tractable decision support in Operations Management (McCloskey, LaCasse, and Cox 2024). For example, Singh et al. (2024a) developed the Interactive Attention Graph Convolution Network, i.e., a model designed to enhance aspect-level sentiment analysis, integrating a dynamic weighting layer with bidirectional memory to capture context effectively. Our work differentiates from extant approaches by postulating a two-step process: (i) retrieving context summaries from a corpus and selectively reviewing them for human intervention and feedback, and (ii) using the generated summaries as the context to the articles' sentiment while the feedback informs the prompts and the LLM generates sentiment data. The importance of context summaries and "summary of summaries", which are central to the CESCSS framework, lies in informing the LLM's estimation and the opportunities they create to assess and interpret the outcome metrics. The "summary of summaries" involves taking context summaries for each article in a corpus and creating a single meta-summary (Mayande et al. 2024).

Our CESCSS framework recognises that extant LLM technologies may generate hallucinations to an extent (Yilmaz, Yilmaz, and Ceylan 2023). Hence, we offer reliability and interpretability by incorporating robust mechanisms that serve as a guardrail while interpreting the context associated with sentiments as a metric. In

essence, our research echoes the necessity for inherently explainable models in decision-making (Rudin 2019) and develops an interpretable LLM-based assessment framework that takes steps toward transparency. Henceforth, this article uses the terms 'interpretable' and 'explainable' interchangeably.

The rest of this study is structured as follows: Section 2 provides the literature background on current approaches to investigating context and sentiment extraction from textual data. Section 3 presents the research context underpinning this research. Section 4 outlines the CESCSS framework. Section 5 presents the results. Section 6 discusses theory and practice implications while highlighting the proposed research approach's limitations and the need for further work. The paper concludes in Section 7.

2. Literature

Figure 1 outlines our proposed context extraction and sentiment interpretation framework, termed CESCSS, to analyse textual data across critical mineral SC echelons. The direct LLM approach (top of Figure 1) feeds text-based articles directly into an LLM along with a prompt that seeks sentiment as the outcome metric for each article. Relying exclusively on such a typical 'black box' approach may lead to convincing yet inaccurate sentiment interpretations, as LLMs can produce outputs that lack transparent reasoning (Ajwani et al. 2024). Our postulated approach (bottom of Figure 1) differs from the direct approach in that we bring in context to pre-process the data before estimating the sentiments (involving nine implementation steps – these nine steps are further explicated in Section 4, Figure 3). Context allows human intervention by interpreting the initial LLM outcome and making suitable corrections in the prompts. To set up our proposed CESCSS framework, we draw on four disparate sets of literature: (i) context generation and interpretation, (ii) sentiment analysis, (iii) sentiment shifts, and (iv) information asymmetry.

2.1. Context generation and interpretation

The importance of summaries while assessing sentiments is recognised (Mayande et al. 2024; Yang et al. 2019). Context identification seeks to summarise the circumstances surrounding SC events or decisions based on patterns in language use within a news article. The extant scientific literature has scrutinised context summarisation from textual information, mainly in the healthcare sector. Indicatively, Sallam et al. (2023) studied the use of ChatGPT in healthcare education by analysing factors

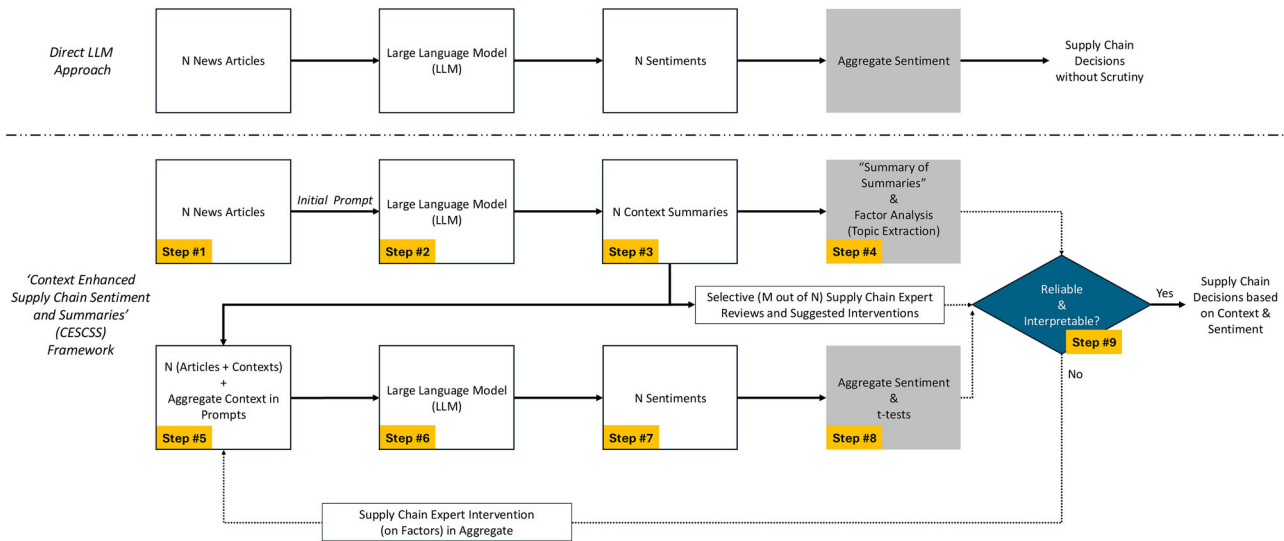


Figure 1. Direct sentiment analysis vs. postulated CESCSS framework.

influencing its acceptance, including perceived usefulness, ease of use, risks, and factors that shape attitudes toward adopting ChatGPT. Specifically, the assessment framework of Sallam et al. (2023) revealed that accuracy, bias, and ethical issues persist, notwithstanding the promising benefits of AI systems powered by LLMs (e.g., personalised learning and improved communication).

Furthermore, Cuberek, Janíková, and Dygrýn (2021) focused on adapting and validating the Physical Activity Questionnaire for Older Children, which was used in a national survey by the Czech Ministry of Education, Youth, and Sports to assess children's physical activity. Differences in the survey questionnaire were investigated via a t-test, while Cronbach's alpha was used to evaluate consistency. Findings indicated the need for extensive data samples to ensure the validity and reliability of the results. Additionally, in business management, Wolfe and Mitra (2024) explored the impact of GenAI on fact-checking organisations to improve response times to misinformation, increase audience engagement, and boost productivity. However, technical skill gaps, value tensions, and concerns over AI's reliability and alignment with the core values of transparency and human expertise limit its adoption. Hence, Wolfe and Mitra (2024) highlighted the need for careful integration of AI in business processes, proposing a design framework centering on content verification, mapping out tensions between AI and fact-checking values, and suggesting a research agenda for responsible AI use in this field.

Regarding context generation, the rise of LLMs enables machines to generate, comprehend, and analyse natural language with unprecedented accuracy. In fields such as healthcare, education, and media, the ability to

predict outcomes, trends, or insights based on text can improve the effectiveness of various applications. Practically, a text structure (e.g., a sentence) is considered, and one can decide where to focus attention to process a large corpus (Vaswani et al. 2017). Paying attention requires context beforehand. Within the body of the LLM itself, the context is provided by feeding data that allows the LLM to adjust weights. In addition, it is also possible to preprocess data (outside the LLM) to help set up the context. For instance, the Retrieval Augmented Generation (RAG) allows machines, secondary data sources, or experts to enhance the prompts and provide context (Li et al. 2022). Recent research focused on developing LLMs with longer context windows, enabling them to store and directly use more information when generating responses (Gadelho 2024). It is worth noting that RAG ideas are beginning to be applied to the SC literature (Gaddala 2023; Ram et al. 2023).

Due to large textual datasets, LLMs are primarily used, thus far, simply for natural language understanding and text generation that are not particularly useful and suffer from hallucinations, i.e., incorrect or baseless explanations (Singh et al. 2024b). Real-world operations require LLMs to generate explanations that are accessible and reliable for high-stakes applications such as decision-making. Hence, research evidence emphasises the importance of providing data-driven explanations of disruptions in areas like SC management for end-to-end decision support to facilitate resilience (e.g., Baryannis et al. 2019; Kosasih et al. 2024). The potential for our work in shaping decision-making for SC resilience is discussed in Section 6, drawing on examples provided by expert informants in critical mineral SCs.

2.2. Sentiment analysis

Sentiment analysis, or opinion mining, is an expanding field of research within Information Processing, focusing on extracting and analysing subjective elements such as opinions, sentiments, attitudes, and emotions from textual information (Varathan, Giachanou, and Crestani 2017). Traditional text retrieval systems primarily handle factual data, but sentiment analysis delves into subjective content, enabling tasks like opinion mining, subjectivity analysis, and emotion detection. With data from sources like social media and blogs increasing, sentiment analysis provides valuable applications, such as improving recommendation systems and aggregating reviews.

Serrano-Guerrero et al. (2015) reviewed fifteen web services offering sentiment analysis functionalities, which are particularly valuable for researchers and developers looking to incorporate sentiment capabilities without developing bespoke algorithms, further reproducing a robust taxonomy of sentiment classification techniques by Medhat, Hassan, and Korashy (2014). Considering the ability to leverage diverse data collections (i.e., long and short texts), such tools can help detect positive, negative, and neutral content, which could apply to SC-related content. Nevertheless, oversimplified classifications (positive, negative, neutral) and difficulty in aggregating sentiment across multiple sentences or paragraphs suggest that techniques could improve precision and functionality. Vishnuthilak et al. (2024) showcased the potential of sentiment analysis in identifying disruptive events and introduced a framework that integrates a trigger-based support system to aid decision-making in SC management, potentially enhancing resilience.

In the healthcare sector, Kumar et al. (2025) analysed Twitter data using an AI-driven approach to address healthcare SC challenges during the COVID-19 pandemic, focusing on issues like shortages of testing kits, oxygen, and hospital beds. The study leveraged natural language processing and machine learning to classify tweets based on need severity and predict locations to facilitate timely crisis response and coordination efforts. However, conducting such analyses using LLMs could be fraught with the risk of hallucinations (Huang et al. 2025). Thus, the analysis process needs to establish reliable interpretations (Liu et al. 2023c). These arguments informed our previously stated RQ#1 – *What reliable LLM-based method can be used to extract context from textual data and enhance the interpretability of SC sentiment analysis?*

2.3. Sentiment shifts

In pharmaceuticals, Ball et al. (2024) investigated the possibility the sentiment of drug reviews on WebMD can

serve as an early-warning indicator of serious drug recalls and whether this relationship varies by the gender of the reviewer. The latter study specifically focused on FDA Class I and Class II recalls associated with severe harm or death and revealed that an increase in negative sentiment was significantly associated with a higher risk of a drug recall. Notably, one standard deviation rise in negative sentiment was linked to a 15.9% increase in recall risk. On the pricing domain, Anderson, Mahar, and Modi (2024) used LLMs to forecast price index changes, where predicted sentiments and allied contextual factors, such as shipping and inventory, served as independent variables. The ability to assess sentiment from linguistic evidence has gained significant attention in natural language processing, especially with the advancement of LLM models (Shamshiri, Ryu, and Park 2024; Zhang et al. 2023). However, context may play a crucial role in shaping sentiment in SC echelons, as the exact words can convey different emotions depending on the surrounding text about pertinent stakeholders' goals or decisions.

Shifts in sentiment help identify and track attitude changes over time, which may reflect evolving conditions within different SC echelons. To this effect, reliability issues regarding the inputs to the sentiment analysis are a crucial challenge for SCs. Evidence from the pharmaceutical sector revealed that female-written reviews provide early warning indicators for serious drug recalls compared to the reviews of males (Ball et al. 2024). Specifically, the respective research results highlighted that in gender-focused operations management, consumer sentiment could be a critical factor in predicting serious product issues, further necessitating a deeper understanding of the role of social media sentiment in healthcare. Furthermore, research on news media and tweets regarding the coronavirus disease 2019 (Omicron phase) and the Ukraine – Russia war (from November 2021 to April 2022) revealed that sentiment and time effect add value to SC risk management. In the energy field, Beyer Díaz et al. (2024) provided a framework combining structured and textual data to predict oil prices and argued for the examination of the impact of adding textual data sources to foster deep learning models' interpretability in the presence of global geopolitical events. Time-varying and shifts in sentiment could help identify unnoticed risks (e.g., logistics tension, ripple effect) (Sadeek and Hanaoka 2023). These arguments influenced our previously articulated RQ#2 – *How can sentiment shifts estimated by an LLM be used to detect disruptions' propagation across SC echelons?*

2.4. Information asymmetry

A constituent that can impact the reliability of sentiment analysis is information asymmetry across SC

stakeholders, which is common in transactions within extended and distributed networks and can be expressed in diverse manners, such as not revealing private information or ‘cheap talking’ (i.e., communicating costless, non-binding, and non-verifiable information). For example, Inderfurth, Sadrieh, and Voigt (2013) examined the impact of information asymmetry on the dyadic relationship in a buyer-supplier setting, demonstrating that information transparency limits out-of-equilibrium behaviours, thus benefiting all the involved SC stakeholders. Furthermore, awareness of market demand holds significant value for both retailers and manufacturers; in online celebrity SCs, research findings indicate that option contracts yield varying payoffs contingent upon the retailer’s capital (Tang, Wu, and Dolgui 2020). In this regard, sentiment analysis could be utilised to quantify trust levels and emotional tone in partner communications, revealing correlations between sentiments and the cooperation observed in SC interactions.

Additionally, Özer, Zheng, and Chen (2011) investigated and modelled the role of trust between suppliers and a manufacturer’s private forecast information in determining capacity investment decisions. In the same vein, Li, Zhang, and Dan (2019) explored the role of information asymmetry in SCs, showing that firms may distort demand signals to maximise profit, thereby influencing competition dynamics and channel cooperation strategies. Such an unequal distribution of information can amplify shifts in sentiment, leading to greater volatility and polarisation as SC stakeholders might interpret incomplete or skewed data differently. Asymmetry and shifts in sentiment can also be attributed to the textual data sources used, as news media mainly focus on the supply side, while social media mirrors the demand side in a SC (Sadeek and Hanaoka 2023). These arguments underpinned our formulation of RQ#3 – *Can LLM-based sentiment analysis reflect shifts based on information asymmetry in SCs?*

3. Research context: cobalt SCs

Cobalt, a critical battery mineral (Zhang, Aydin, and Parker 2022), was employed as a case study to illustrate the proposed CESCSC framework. This study considered three echelons in the cobalt SC: mining (upstream), processing raw material and component manufacturing (midstream), and assembly applications (downstream) (Figure 2).

Upstream the SC, cobalt is extracted in twenty-one countries, with most mining operations concentrated in the Democratic Republic of Congo (DRC), Indonesia, Russia, and Australia. Specifically, we focused on mining operations in the DRC, which supplies around 80%

of the global cobalt demand (Cobalt Institute 2024). For midstream, we focused on China, which processes around 68% of cobalt (Quiggin and King 2023). China is the world’s largest producer of refined cobalt, primarily utilising partially processed cobalt imported from the DRC (U.S. Geological Survey 2024). For downstream, we focused on the U.S.A. and Canada, countries where significant cobalt applications are being designed into products such as electric vehicles (EVs) (Pelon et al. 2021).

Sentiments are a key metric affecting cobalt SC decisions and policymaking. In this view, we surveyed participants at the ‘2024 Cambridge Critical Mineral Supply Chain Resilience Workshop’ held at the University of Cambridge, UK. Informants ($N = 58$) included 42% academics, 17% policymakers, and 41% industry representatives from the critical minerals sector. A list of the informants, along with their roles and salient points, is inserted in Table S1 in the Supplementary Material (Section S1).

Two primary insights emerged from this survey. First, context-driven sentiments are crucial for three types of decision-making: (i) financial, (ii) SC management, and (iii) government policy. Approximately 66% of participants indicated that financial analysts rely on news information and allied shifts in sentiments when making decisions; around 75% of the participants identified SC managers as the primary users of sentiment-driven information, while 81% pointed to policymakers. Second, the participants emphasised that contextualised sentiment data is fundamental to strategic decision-making, confirming their essential role in guiding industry practices. For example, one surveyed participant pointed out that:

Market valuation (particularly on critical mineral commodities) is influenced by the combined effects of commodity pricing and supply chain sentiment. [Informant #24]

We further use the survey findings to discuss decision-making in Section 6. A bibliometric analysis of the informants’ quotes is inserted in Figure S1 in the Supplementary Material (Section S1). The resulting map shows that the sentiment is a central (hub) construct that co-evolves with other factors, such as prices, government policies, and risks and shortages created by SC disruptions.

4. CESCSC framework

In this section, we set out the methodological foundations and implementation steps for the AI-driven CESCSC framework, aiming to identify computational principles and procedures that can drive reliability and interpretability. Consistent with Mayande et al. (2024),

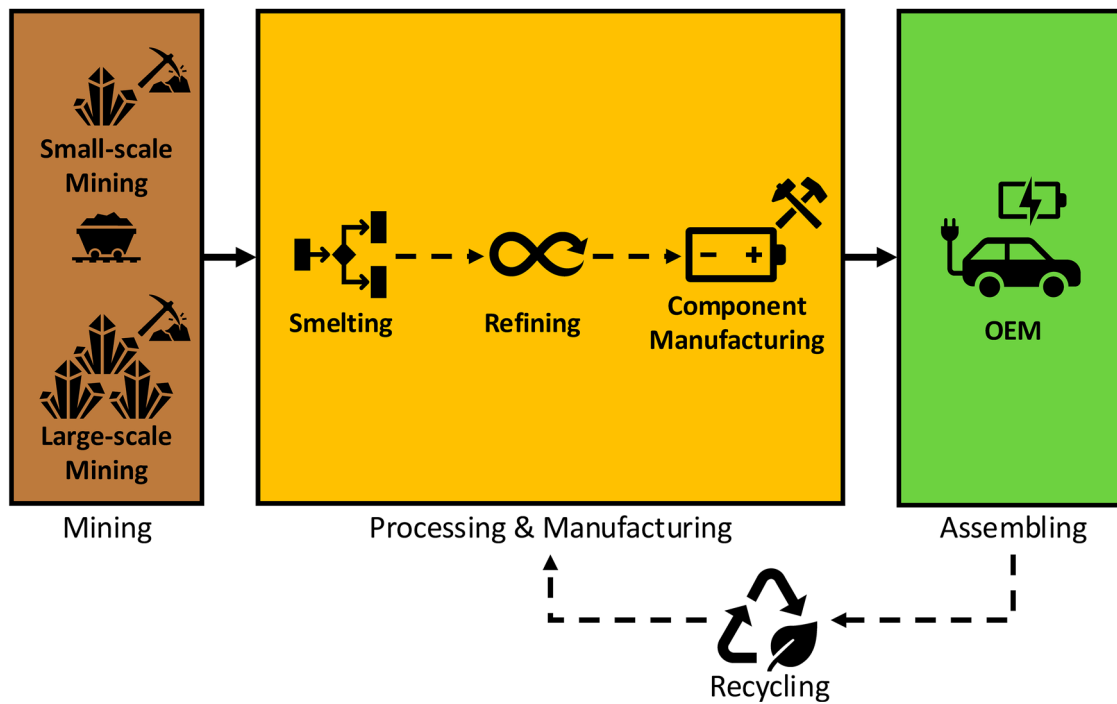


Figure 2. Cobalt SC.

our research approach focuses on input augmentation instead of fine-tuning the LLM. While Mayande et al. (2024) tested their method on marketing data, we applied the underlying ideas to the SC realm. Additionally, our work amounts to a particular case of RAG in that we deploy Reinforcement Learning from Human Feedback (Sun et al. 2023) (as shown in the targeted and aggregate feedback loops at the bottom of Figure 1).

The research stages of our proposed CESCSS framework include: (i) context generation through LLM, (ii) sentiment analysis through LLM, and (iii) reliability and interpretability. Figure 3 captures the research stages, respective implementation steps, and outcomes of the CESCSS framework. For clarity, feedback is excluded from this visualisation. Figure S2 in the Supplementary Material (Section S2) depicts a more detailed methodology process flowchart.

The first research stage involved defining the analysis scope, compiling relevant textual data from news articles, and developing the corpus with LLM-generated contextual information. We prompted contextual information by summarising key aspects of each news article related to the cobalt SC. Our research argues that textual data summaries can provide a compelling context for SC sentiment analysis.

In the second stage of this research, we conducted sentiment analysis at the document level using an LLM, utilising the summaries generated in the first stage to provide

contextual information. We also performed the analysis without contextual information to gain a comparative view of how context affects sentiment interpretation. Furthermore, we conducted statistical t-tests comparing sentiment results across scenarios regarding the country of occurrence and timelines of the disruptive event to ensure rigour. The latter tests helped determine if the LLM-generated summaries significantly altered sentiment outcomes compared to the uncontextualised analysis. The analysis also revealed insights into the impact of contextual generation on sentiment detection while underscoring LLMs' roles in refining sentiment interpretation in textual data.

In our research's third stage, LLM-generated contexts and sentiment scores were validated by human experts to ensure output reliability and interpretability. We compared the LLM-generated summaries and sentiment scores with those from experts independently evaluating the same news articles. All eight experts contributing to the validation assessment possess at least a Master's degree in relevant areas, such as industrial engineering or manufacturing; several hold doctoral degrees in similar fields. Each expert has experience in fieldwork and data analysis, focusing on decision-making pertaining to critical mineral SCs. Reliability assessment and interpretation underscore the importance of human scrutiny when using LLMs for nuanced textual analysis.

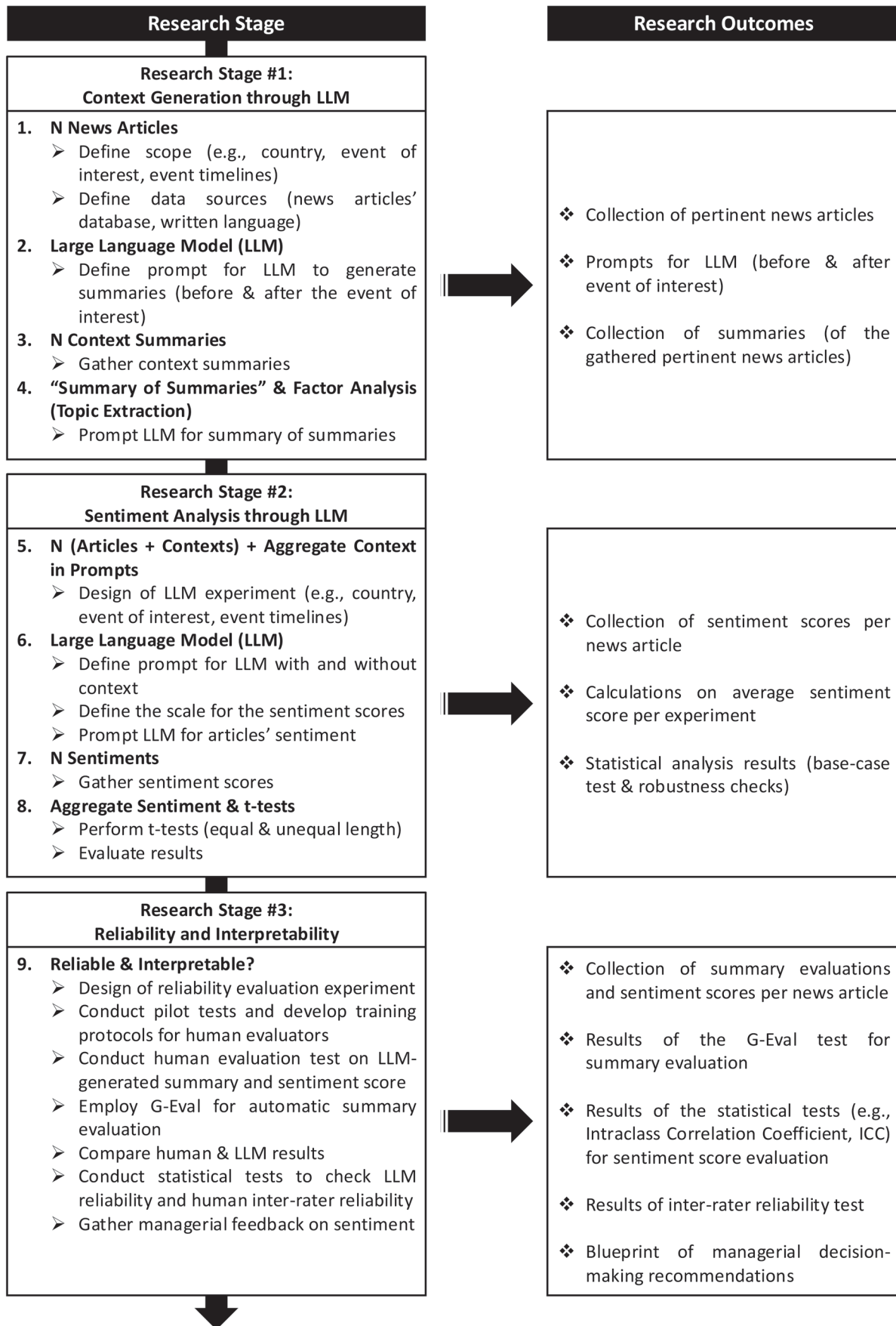


Figure 3. Research stages and implementation steps of the CESC framework [implementation steps 1 through 9 correspond to Figure 1, but feedback loops are not shown for brevity].

For implementation, we employed the Gemini 1.5 Flash model to test our arguments. Gemini 1.5's extensive context window, capable of handling up to one million tokens (Google AI for Developers 2024), enabled us to summarise large text corpora, which is particularly needed in processing and analysing SC-related news data.

4.1. Context generation through LLM (research stage #1)

In this research stage, the primary focus was on collecting relevant news articles within the predefined scope, including specific countries and timelines of a disruptive event. This step ensured that the gathered data reflected the geographical and temporal boundaries necessary for accurate analysis. The news articles were collected automatically, without manual screening, based on the criteria described below. We then deployed an LLM to generate concise summaries for each retrieved article to streamline the process and enhance efficiency. The summaries served as a structured foundation by distilling the essential information and contextual elements from the news articles, providing a coherent and organised base supporting the subsequent analysis stages. Specifically, by using LLMs, we aimed to reduce the complexity of data processing while ensuring that the extracted information remained accurate and relevant, which is crucial for the integrity and reliability of the overall analysis.

We first defined the data collection scope by focusing on specific countries and a disruptive event in the cobalt SC. Specifically, we selected the DRC (upstream extraction), China (midstream refining), and the U.S.A. and Canada (downstream consumption) based on their roles in the SC. The 2023 truck strike in the DRC was chosen as the event of focus due to its significant disruption to global cobalt trade flows (Luk and Njini 2023), with a 12-month analysis period, i.e., six months before and six months after the strike. This timeframe allowed for assessing pre- and post-disruption SC dynamics across various operations echelons and regions.

After that, we retrieved secondary data for our analysis by leveraging the Factiva international news monitoring and business intelligence platform (<https://www.dowjones.com/professional/factiva/>) to ensure extensive news coverage regarding the cobalt SC, along with transparency in our data collection process. Factiva is an international news database offering access to over 33,000 premium publications, including major international newspapers, industry-specific trade journals, and newswires (Dow Jones 2024). Furthermore, Factiva is frequently employed in academic and industry research for SC

Table 1. Summary statistics of the collected news articles.

Event time frame	Before		After	
	Sample size (articles)	Average word count per article	Sample size (articles)	Average word count per article
Country/region				
DRC	136	1,248	127	1,249
China	735	2,652	538	3,129
U.S.A. and Canada	1,963	8,394	1,669	13,619
Articles & Ave.	2,834	6,562	2,334	10,528
Word Count				
Total Word Count	43,168,304 (across 5,168 Articles)			

analysis, providing reliable and extensive business intelligence data (Geng, Lam, and Stevenson 2022; Son, Roscoe, and Sodhi 2025).

To ensure methodological rigour, we adopted a structured selection approach inspired by Pratelli et al. (2024), who emphasised representativeness and transparent article sampling in LLM-based evaluations of journalistic content. In particular, we selected every news article available in the specified database, filtered by the region of 'Democratic Republic of Congo', 'China', 'U.S.A. and Canada', and the disruptive event 'truck strike', within a 12-month analysis period (i.e., six months before and six months after the strike). For comparative analysis, we employed the search keywords 'cobalt' and 'supply chain*' distinctively between the periods before and after the disruptive event. Filters were utilised to refine the results by language (i.e., English) and country, ensuring geographic relevance to the identified SC echelons (i.e., DRC, China, U.S.A. and Canada). There was no selection bias following the completion of the filtering process. Table 1 presents the summary statistics of the collected news articles.

Table A1 in Appendix 1 illustrates two prompt templates for generating context summaries of the gathered news articles utilising the Gemini 1.5 Flash model, explicitly focusing on cobalt and SCs. The pre-strike template solicits summaries centred on cobalt and SCs, whereas the post-strike template broadens the scope to contain cobalt, SCs, and/or truck strikes. This approach facilitated capturing shifts in context or sentiment resulting from the truck strike's effect on the SC. The collected news articles were collated in '.rtf' format. We then split the news articles into individual ones using Python's regular expression matching library. A BigQuery table with the retrieved news articles and their metadata was created from which any LLM could read the corpora.

4.2. Sentiment analysis through LLM (research stage #2)

We set out to conduct sentiment analysis to evaluate the sentiment across various operations echelons of the

Table 2. Prompt template for generating sentiment.

Prompt template for analysing news sentiment within a given context using Gemini 1.5 Flash

Given the context: {context}, analyse the sentiment of the following texts: "{article}".

Provide a sentiment score ranging from -1 to 1 (inclusive). Please only print the sentiment score as a numeric value within the range of -1 to 1 . A value ranging from 0.5 to 1 indicates very positive sentiment, 0.1 to 0.5 indicates positive sentiment, -0.1 to 0.1 indicates neutral sentiment, -0.1 to -0.5 indicates negative sentiment, and -0.5 to -1 indicates very negative sentiment.

Sentiment:

cobalt SC before and after the 2023 truck strike. By scrutinising the sentiment in the identified regions of respective operations over the specified timeframe, we ascertained the extent to which the disruptive event influenced the overall sentiment within each SC echelon.

The sentiment analysis was conducted at a document level, examining the entire news article to determine its overall polarity. The latter approach allowed us to capture broad sentiment trends within each region's news coverage, providing insights into how the cobalt SC was perceived before and after the truck strike. Considering the substantial volume of textual data and the significant number of API calls necessary for context generation and sentiment analysis, we utilised LangChains and readily available parallel computing libraries to distribute tasks across multiple CPUs. To further optimise these parallel tasks, we deployed the code on Google Compute Engine virtual machines equipped with four virtual CPUs and 16GB of memory (i.e., e2-standard-4). The dependence on parallel computing libraries and virtual machines is essential for scaling these technical capabilities, emphasising optimised time costs and memory utilisation (Devlin et al. 2018), which is paramount given the extensive corpus of news articles about multi-echelon SCs.

4.2.1. Prompting for sentiment

In the same vein as context generation, we utilised the Gemini 1.5 Flash to generate sentiment scores for the summaries of the retrieved news articles (thirty runs per summary article). We prompted a numerical sentiment score for multiple runs to ensure the robustness and reliability of sentiment analysis for news articles. This approach mitigated variability and outliers by averaging results from several prompts, providing a more consistent overall sentiment assessment using LLMs (Chohan 2016). Table 2 presents the prompt template for analysing news articles' sentiment within the given context (i.e., before and after the truck strike event on the cobalt SC).

4.2.2. Design of experiments

Considering the disruptions associated with a truck strike and the imperative to maintain SC operations performance standards, informed decision-making is pivotal to guiding resilience responses (Faruquee, Paulraj, and Irawan 2023). To this effect, following the methodology outlined by Ivanov (2022), we structured our experiments to investigate news articles' sentiments before and after a disruptive event. Table 3 inserts the Design of Experiments. We identified three base test sets with the intent to assess RQ#1, RQ#2, and RQ#3, respectively, along with six robustness checks.

4.3. Reliability and interpretability (research stage #3)

This study assessed two fundamental outputs of our methodological approach through human expert evaluations. First, we evaluated the quality of the news articles' summaries utilised as contextual input for sentiment analysis. Second, we assessed the reliability of the sentiment scores assigned to news articles (see Appendix 2 for further details on the process). We distributed forms to collect SC expert evaluation results, with training provided to each expert before the evaluation to explain the questions in the forms. The survey was piloted with a SC expert to refine its structure, ensuring clarity and alignment with the study's objectives. The evaluation phase involved eight participants with SC knowledge and expertise providing assessments on predefined factors. Results were summarised, showing participant numbers, answered questions, and scores for summary quality and sentiment, facilitating a thorough reliability analysis of the LLM-generated content. This dual validation process, covering summary quality and sentiment accuracy, strengthened the robustness of our methodology by integrating automated and expert-based evaluations.

4.3.1. Summary evaluations

This study employed the G-Eval method (Liu et al. 2023a) to assess the quality of the LLM-generated summaries. This approach uses GPT-4 (i.e., an advanced AI language model) to automatically evaluate coherence, consistency, fluency, and relevance in summaries produced through natural language generation. Unlike traditional lexicon-based approaches, G-Eval evaluates summaries by analysing their coherence, consistency, fluency, and relevance by applying a Chain-of-Thought and form-filling paradigm, where the evaluation steps are driven by comprehensive intermediate steps produced by the LLM. This approach demonstrated an improved correlation with human evaluations in assessing natural language generation outputs. In line with the G-Eval study,

Table 3. LLM design of experiments.

Scenario ID	Intent	Scenario	Context			Time Frame		Country/Region			Window (Days)	Sample 1 Size	Sample 2 Size	t-test Type
			Article Summary	Aggregated Summary	No Context	Before	After	China	DRC	U.S.A. and Canada				
Base experiment *														
1	RQ#1	Sentiment with and without context	1		1	1	1	1	1	1	±180	5,168	5,168	Paired
2		Sentiment with context before and after strike	1			1	1	1	1	1	±180	2,834	2,334	Welch's
3		Sentiment without context before and after strike			1	1	1	1	1	1	±180	2,834	2,334	Welch's
4		Sentiment with and without context after strike	1		1		1	1	1	1	±180	2,334	2,334	Paired
5		Sentiment with and without context before strike	1		1	1		1	1	1	±180	2,834	2,834	Paired
Robustness test *														
6	Robustness tests for RQ#1	Sentiment with context using article summary and aggregated summary	1	1		1	1	1	1	1	±180	5,168	5,168	Paired
7		Sentiment without context and using aggregated summary as context		1	1	1	1	1	1	1	±180	5,168	5,168	Paired
8		Sentiment with and without context with ±30 days' time window	1		1	1	1	1	1	1	±30	925	925	Paired
9		Sentiment with and without context with ±7 days' time window	1		1	1	1	1	1	1	±7	10	10	Paired
10		Sentiment with context before and after strike in China	1			1	1	1			±180	735	538	Welch's
11	RQ#2 and RQ#3	Sentiment with context before and after strike in DRC	1			1	1		1		±180	136	127	Welch's
12		Sentiment with context before and after strike in U.S.A. and Canada	1			1	1			1	±180	1,963	1,669	Welch's
13		Sentiment with context before strike between China and DRC	1			1		1	1		±180	735	136	Welch's
14		Sentiment with context before strike between China and U.S.A. and Canada	1			1		1		1	±180	735	1,963	Welch's
15		Sentiment with context before strike between DRC and U.S.A. and Canada	1			1			1	1	±180	136	1,963	Welch's
16		Sentiment with context after strike between China and DRC	1				1	1	1		±180	538	127	Welch's
17		Sentiment with context after strike between China and U.S.A. and Canada	1				1	1		1	±180	538	1,669	Welch's
18		Sentiment with context after strike between DRC and U.S.A. and Canada	1				1		1	1	±180	127	1,669	Welch's

*1: The factor has the indicated value.

we prompted GPT-4 twenty times per article for each dimension to calculate an average result. We then analysed the mean and variance of the output scores to assess the effectiveness of these summaries across four dimensions: (i) coherence, (ii) consistency, (iii) fluency, and (iv) relevance. By focusing on content rather than predefined lexicons, we seek to illustrate how context influences summary quality and emphasise G-Eval's potential to improve conventional evaluation methods in summarisation tasks.

Concurrently, we undertook an expert evaluation in which human evaluators assessed the precise dimensions. We supplied eight experts with an identical set of sample summaries along with the evaluation guidelines utilised in G-Eval for human assessment. Thereafter, we evaluated the correspondence between the automated assessments and human judgments by calculating and comparing their respective means, variances, and 95% confidence intervals.

4.3.2. Sentiment score evaluation

This study evaluates the reliability of the LLM-generated sentiment score. A key challenge in sentiment analysis is handling ambiguity and differences in sentiment perception, particularly when human judgment is involved. Therefore, sentiment analysis must also account for the variability in how humans interpret sentiment (Soleymani et al. 2017). Consequently, we performed factor analysis (topic extraction) to identify key dimensions in the sampled articles, providing a reference to guide human evaluators in assessing sentiments. Similar approaches were adopted by several prior works, like Ozyurt and Akcayol (2021), to extract aspects for sentiment analysis. We applied two methods: (i) Principal Component Analysis to reduce data dimensionality and identify significant factors, and (ii) Latent Dirichlet Allocation for topic modelling to highlight the main themes in the news articles (Karl, Wisnowski, and Rushing 2015). Through this process, we identified six relevant factors in the cobalt SC: (i) Cobalt, (ii) Battery, (iii) EVs, (iv) Market, (v) Supply Chain (and/or truck strike events for news published after the strike event), and (vi) Country.

Similarly, a sentiment score evaluation form was distributed to the same group of SC experts with the original articles of the six sampled news articles and the corresponding LLM-generated sentiment scores. Experts were first asked to independently evaluate the sentiments of each of the six factors. They were provided with three options: negative, neutral, and positive; which were then converted into numerical sentiment scores for further analysis. The experts were then asked to give an overall

sentiment score for each article, considering all the factors.

Subsequently, we employed statistical methods to assess the reliability of the data gathered. We first compared sentiment scores generated by humans and those produced by LLM to evaluate their level of agreement and consistency. This comparison involved calculating the intraclass and Pearson correlation coefficients (Lahat et al. 2023). Additionally, we applied a mixed-effects linear regression model to examine how the type of rater (i.e., human or LLM) influenced the sentiment scores (Bakker et al. 2022). To ensure comparability, we randomly selected four out of the thirty sentiment scores generated by the LLM to match the number of human raters.

Finally, to account for variability in human perception of news sentiment, we assessed inter-rater reliability using Cronbach's alpha, ensuring consistency across expert ratings to strengthen the reliability of our findings (Taber 2018). This measure evaluates internal consistency, indicating how well items reflect the same construct. We calculated Cronbach's alpha across seven items in the form, including six identified factors and the overall article sentiment score, to confirm that expert evaluations consistently reflected each article's sentiment within its context, thereby enhancing the validity of our research. Similarly, a Cronbach's alpha score was calculated to assess the internal consistency of the outputs generated by the LLM.

4.3.3. Aggregate managerial feedback

This study also developed a "summary of summaries" approach to facilitate informed managerial decision-making for SC resilience. For each timeframe (i.e., before and after) and each country (i.e., DRC, China, U.S.A. and Canada), we used Gemini 1.5 Flash to generate an overall summary of all individual summaries within the specified timeframe and country. To assess the effectiveness of this approach, four SC experts were tasked with evaluating the "summary of summaries" for the three countries and the two different specified timeframes, i.e., six "summary of summaries" in total. The evaluation process was structured into two main stages of paired comparisons. In the first stage, experts conducted a within-country comparison, examining pre – and post-event "summary of summaries" to identify changes in tone or content that could influence context-driven decisions (i.e., financial, SC, policy). In the second stage, the involved experts performed a cross-country comparison of post-event summaries, using colour-coded annotations to capture differences across countries. The highlighted differences in summaries from each expert were then consolidated into

a summary table to facilitate comparison and analysis. Section S3 in the Supplementary Material documents all six "summary of summaries" with colour codes highlighting differences in interpretations.

5. Results

This section delineates the findings derived from our analysis, organised into four principal subsections. Section 5.1 elaborates on the testing outcomes of our RQs concerning the LLM sentiment scores whilst scrutinising the variability of sentiment across diverse contexts and events in Section 5.2. Section 5.3 presents a comparative analysis between human evaluations and scores generated by the LLM, underscoring the similarities and differences observed in rating patterns and reliability across various metrics. Lastly, Section 5.4 interprets the "summary of summaries", which integrates qualitative insights tailored for a managerial audience and conducts both within-country and cross-country comparisons to evaluate sentiment fluctuations preceding and following the disruptive event under consideration.

5.1. LLM sentiment score: t-test results

The results presented in Table 4 provide an overview of a series of t-tests performed on the scenarios within our Design of Experiments (Table 3) to evaluate sentiment scores across varying conditions, particularly contextual factors and the timelines of the truck strike event in the DRC. Each scenario comprises two samples, accompanied by their respective means (allowing for unequal variances); the t-statistic and p -value indicate the statistical significance of the differences observed between the two samples.

The base experiment scenarios (i.e., Scenarios 1 through 5 and Scenarios 12 through 18) address three sets of test results for our RQs. These tests show significant differences across each paired t-test for averages. Scenario 1 demonstrates a highly significant difference ($p = 6.203\text{E-}28$), with a negative t-statistic of -11.02 , reflecting that sentiment scores tend to be lower without context. Similarly, Scenarios 2, 3, 4, and 5 exhibit significant differences, as evidenced by p -value below 0.05, illustrating variations in sentiment before and after the truck strike event and with or without contextual elements.

The robustness test scenarios for RQ#1 (i.e., Scenarios 6 through 11) further examine the variances in sentiment within different contexts across regions and temporal frames. Scenario 6 analyses sentiment through various summary types and reveals a statistically significant positive difference ($t = 4.05$, $p = 5.32\text{E-}05$), implying that the type of summary utilised influences sentiment

scores. Scenario 7 illustrates a notable distinction in sentiment when contextual information is either included or omitted, using an aggregated summary ($t = -11.57$, $p = 1.48\text{E-}30$), with a positive difference suggesting that sentiment is amplified when the context is aggregated. Conversely, Scenario 9, which compares sentiment over a ± 7 -day period, does not produce significant results ($p = 0.28$), indicating that a short-term timeframe does not markedly impact sentiment perception; perhaps this is an artifact of the small sample size in this instance.

Finally, the regional comparisons illustrate notable sentiment dynamics across countries before and after the truck strike. For instance, Scenario 14 (before the strike, China vs. U.S.A. and Canada) and Scenario 15 (before the strike, DRC vs. U.S.A. and Canada) demonstrate highly significant differences with p -value of $9.85\text{E-}37$ and $1.42\text{E-}08$, respectively, alongside substantial positive differences ($\Delta = 0.22$ and $\Delta = 0.25$, respectively). This indicates considerable regional variations in sentiment, with the U.S.A. and Canada exhibiting higher sentiment scores than China and the DRC. Comparable patterns emerge post-strike, with Scenario 17 (China vs. U.S.A. and Canada) and Scenario 18 (DRC vs. U.S.A. and Canada) revealing significant positive differences. The latter suggests that sentiment in the U.S.A. and Canada remains consistently elevated. These findings underscore the critical role of contextual and regional factors in influencing sentiment responses across various scenarios.

5.2. Assessment of summaries

Figure 4 presents the results of the summary evaluations using G-Eval with GPT-4, along with comparable assessments made by human evaluators across four dimensions of the summary readability and alignment with the original content, i.e., coherence, consistency, fluency, and relevance (Liu et al. 2023a). Both average human-rated and GPT-4 assigned scores range between 4 and 5, mostly above 4.5, across all four dimensions for the sampled articles. Upon examining each dimension, we observed that consistency received a relatively low score of 4.33 from experts, accompanied by the most considerable variance of 0.92, indicating potential outliers in human evaluations. In contrast, fluency achieved the highest scores according to both human and GPT-4 assessments. Since human evaluators and GPT-4 were restricted to integer scores on a 1–5 scale, the results suggest that the Gemini-generated summaries are of high quality. Even when considering variability among raters, as indicated by the variance and confidence intervals, these summaries appear robust to serve as the context for sentiment scoring.

In addition, the scores indicate that GPT-4 marginally but consistently rated the summaries higher than human

Table 4. LLM Experiment results (t-tests).

Scenario ID	Intent	Scenario	t-statistic	p-value	Sample 1			Sample 2			Difference
					Size n_1	Mean μ_1	Var σ_1^2	Size n_2	Mean μ_2	Var σ_2^2	$\Delta = \mu_2 - \mu_1$
Base experiment											
1	RQ#1	RQ1: Sentiment with and without context	−11.02	6.20E-28**	155,040	0.26	0.17	155,040	0.24	0.16	−0.03
2		RQ2: Sentiment with context before and after strike	2.60	0.01**	85,020	0.28	0.16	70,020	0.25	0.19	−0.03
3		RQ3a: Sentiment without context before and after strike	2.16	0.03*	85,020	0.25	0.15	70,020	0.22	0.16	−0.02
4		RQ3b: Sentiment with and without context after strike	−8.09	9.45E-16**	70,020	0.25	0.19	70,020	0.22	0.16	−0.02
5		RQ3c: Sentiment with and without context before strike	−7.92	3.37E-15**	85,020	0.28	0.16	85,020	0.25	0.15	−0.03
Robustness test											
6	Robustness tests for RQ#1	Sentiment with context using article summary and aggregated summary	4.05	5.32E-05**	155,040	0.26	0.17	155,040	0.25	0.17	−0.01
7		Sentiment without context and using aggregated summary as context	−11.57	1.48E-30**	155,040	0.24	0.16	155,040	0.25	0.17	0.02
8		Sentiment with and without context with ± 30 days' time window	−3.40	7E-4**	27,750	0.23	0.16	27,750	0.25	0.15	0.03
9		Sentiment with and without context with ± 7 days' time window	−1.15	0.28	300	0.13	0.42	300	0.20	0.41	0.08
10		Sentiment with context before and after strike in China	0.81	0.42	22,050	0.12	0.16	16,140	0.11	0.18	−0.02
11	RQ#2 and RQ#3	Sentiment with context before and after strike in DRC	2.13	0.03*	4,080	0.10	0.22	3,810	−0.03	0.29	−0.13
12		Sentiment with context before and after strike in U.S.A. and Canada	2.55	0.01*	58,890	0.35	0.14	50,070	0.31	0.16	−0.03
13		Sentiment with context before strike between China and DRC	0.54	0.59	22,050	0.12	0.16	4,080	0.10	0.22	−0.02
14		Sentiment with context before strike between China and U.S.A. and Canada	−13.09	9.85E-37**	22,050	0.12	0.16	58,890	0.35	0.14	0.22
15		Sentiment with context before strike between DRC and U.S.A. and Canada	−6.01	1.42E-08**	4,080	0.10	0.22	58,890	0.35	0.14	0.25
16		Sentiment with context after strike between China and DRC	2.69	0.01**	16,140	0.11	0.18	3,810	−0.03	0.29	−0.14
17		Sentiment with context after strike between China and U.S.A. and Canada	−10.06	1.37E-22**	16,140	0.11	0.18	50,070	0.31	0.16	0.21
18		Sentiment with context after strike between DRC and U.S.A. and Canada	−7.16	4.50E-11**	3,810	−0.03	0.29	50,070	0.31	0.16	0.35

* $p < 0.05$; ** $p < 0.01$.

§ The sample sizes in Table 4 are 30 times the sample sizes in Table 3 because each assessment was made 30 times.

Human vs LLM Mean Summary Evaluation Score with 95% Confidence Intervals

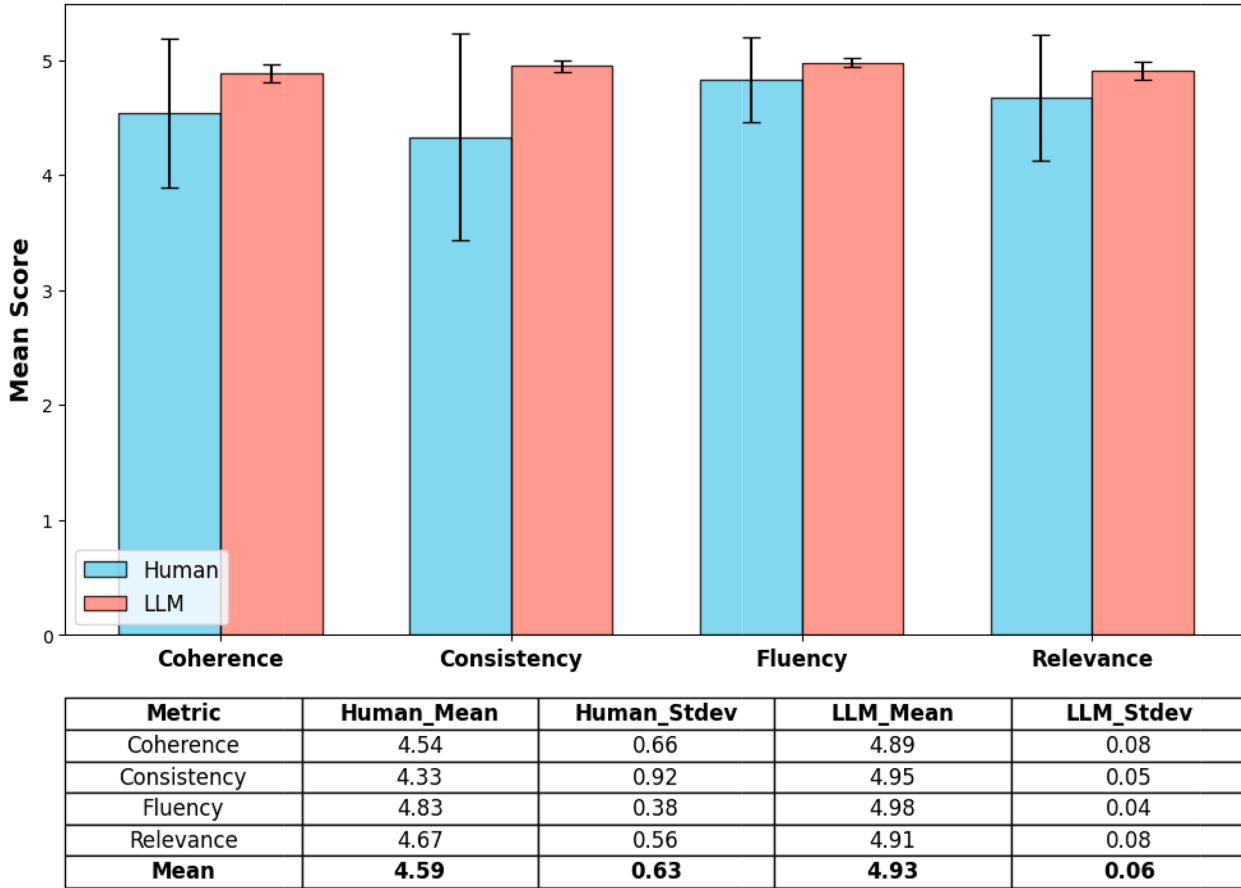


Figure 4. Comparison of human and LLM evaluations of LLM-generated summaries across coherence, consistency, fluency, and relevance.

evaluators across all dimensions. This finding is consistent with the original G-Eval study (Liu et al. 2023a). Regarding the variance, the assessments made by GPT-4 across the four dimensions exhibit minimal variance, denoting high inter-rater reliability. There is a more significant variability among human raters compared to the output generated by GPT-4. This variability is anticipated, given the subjective nature of human evaluations and the small size of the human raters' cohort. Moreover, the 95% confidence interval for human ratings fully covers the confidence interval for the GPT-4 ratings despite the relatively small sample size, suggesting that human raters and GPT-4 ratings generally agree. The relatively large variability in human assessments and the high mean human evaluation scores resulted in confidence intervals whose upper bounds frequently exceeded the maximum possible rating of 5 on the measurement scale. Section S4 in the Supplementary Material provides an explicit comparison with current approaches, highlighting the advancements in our study.

5.3. Assessment of sentiment scores

Figure 5 illustrates the distribution of sentiment scores assessed by the LLM and human experts for the six sampled news articles. Overall, the sentiment scores from human experts and the LLM demonstrate a similar trend across each article, despite the scores from human experts showing greater variability.

To assess the answer to RQ#1 around reliability, we employed a series of statistical methods to evaluate the reliability of LLM-generated sentiment scores. This process involved comparing the sentiment scores produced by human experts and the LLM, then examining the internal consistency within the results from both human evaluators and the LLM. First, we employed the Intraclass Correlation Coefficient (ICC), which measures the level of agreement or consistency among raters (Müller and Büttner 1994). In this study, we used ICC to assess score consistency across multiple raters, including human and LLM raters for each sampled article group. For each sampled news article, we included four sentiment scores from human experts and four from the LLM, randomly

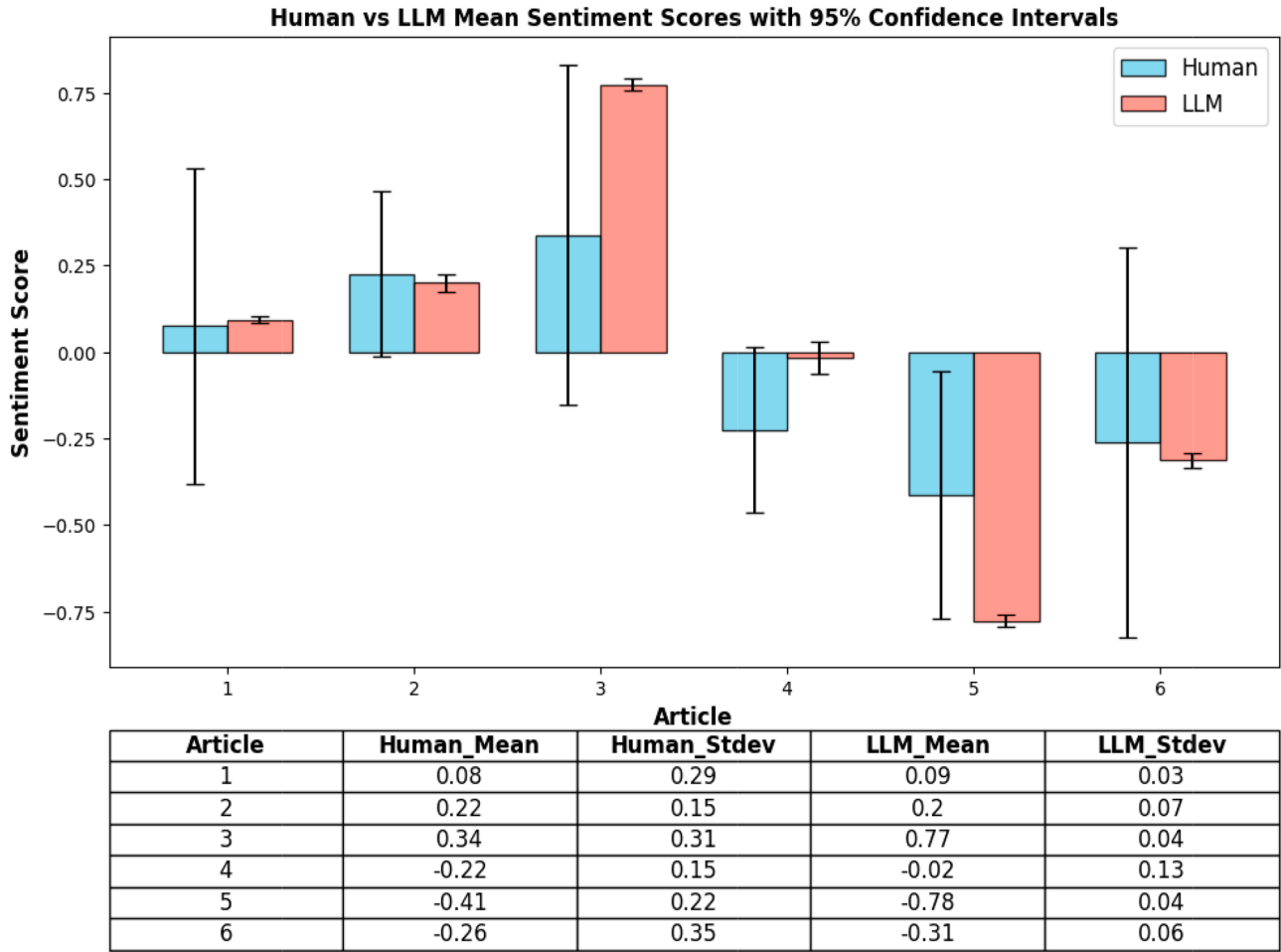


Figure 5. Comparison of human and LLM sentiment scores (for six sampled news articles).

Table 5. Intraclass correlation coefficient.

Type	ICC	CI 95%	F	p-value
ICC(1, k)	0.962	[0.89, 0.99]	26.240	6.377E-12
ICC(2, k)	0.962	[0.89, 0.99]	26.628	5.187E-11
ICC(3, k)	0.962	[0.89, 0.99]	26.628	5.187E-11

selected from thirty generated scores. ICC estimates and their 95% confidence intervals were calculated using Python Pingouin. Table 5 presents the ICC scores. We obtained an ICC score of around 0.96, indicating excellent agreement among raters. At the same time, the low p -value indicates that the agreement observed is statistically significant, meaning that the differences between human and LLM sentiment analysis results are not due to random chance.

Using the same evaluation data, a Pearson correlation coefficient of 0.916 was calculated to assess the linear relationship between sentiment scores from human experts and those generated by the LLM. This strong positive correlation indicates that as human evaluators' sentiment scores increase, the LLM-generated scores also rise, reflecting a high level of agreement between the

two assessments. This robust correlation further confirms that the LLM's sentiment analysis closely aligns with human judgments, underscoring its reliability as a tool for sentiment evaluation.

We also implemented a mixed-effects model to compare sentiment scores provided by human raters and LLM. We mixed the scores from humans and LLM and labelled them with the corresponding rater types. A mixed-effects linear regression model was fitted using the formula $SentimentScore \sim RaterType$, with the grouping factor set to the articles. This allowed us to assess the impact of rater type (fixed effect) on sentiment scores while accounting for the variability associated with different articles (random effect). Table 6 presents the results of the fitted mixed-effects linear regression model. Overfitting and model fit metrics were not the primary concerns in this case, as the model's role is to reveal if there is a systematic difference between human and LLM scores. The mixed-effects model results indicated no statistically significant effect of rater type on sentiment scores, as reflected by the coefficient for rater type (0.048) with a p -value of 0.444. This coefficient value also means LLM

Table 6. Mixed-effects model results.

	Coefficient	Std Err	Z	$P > z $	CI 95%
Intercept	−0.044	0.151	−0.291	0.771	[−0.339, 0.251]
RaterType[T.LLM] ^a	0.048	0.063	0.766	0.444	[−0.075, 0.171]
Group Var ^b	0.124	0.371			

^aRepresents the effect of the LLM as compared to the reference category (human) in the RaterType variable.

^bRepresents the estimated variance attributed to the random effect for Article, which is the grouping variable.

Table 7. Inter-rater reliability for expert sentiment evaluation.

Metric	Value
Mean of item mean	−0.060
Mean of item variance	0.422
Mean of inter-item covariance	0.133
Cronbach's alpha	0.764

tends to rate sentiment slightly higher than humans (by 0.048 points on average, which is 2.4%), but this difference is not statistically significant. This suggests that the sentiment scores generated by the LLM do not differ significantly from those assigned by human raters. The group variance of 0.124 indicates variability among articles but does not contribute to a significant difference in sentiment scores based on rater type. Overall, these results imply that the LLM's sentiment assessments are comparable to those of human evaluators, demonstrating reliability in the LLM's performance in this context-driven sentiment analysis approach.

We further assessed the inter-reliability within human and LLM ratings separately. Table 7 presents the inter-rater reliability assessment of expert responses in evaluating sentiment scores measured through Cronbach's alpha. This metric was calculated across all survey items related to sentiment scores, including six factors (cobalt, batteries, EVs, market conditions, supply chain, and country) and the overall sentiment score included in the survey. The reliability coefficient, Cronbach's alpha (0.764), signifies an acceptable level of internal consistency within this human expert sentiment evaluation, surpassing the 0.7 threshold typically required in social sciences for reliable measurement.

Additionally, the mean inter-item covariance of 0.133 suggests a moderate coherence in sentiment among items, indicating that experts' ratings on one item exhibit a degree of commonality with ratings on others, thereby facilitating an integrative analysis of sentiment patterns across the evaluation's domains. Similarly, we calculated Cronbach's alpha for the LLM-generated sentiment scores on the same sampled articles. With thirty runs per article, Cronbach's alpha was 0.999, and when using a randomly selected subset of four runs, the value was 0.996, both indicating an acceptable level of internal consistency. The calculated Cronbach's alpha values suggest

that human raters and LLM consistently analyse news sentiment within their groups.

5.4. Interpretability of "summary of summaries"

To enhance the manageability and actionability of the extensive data accessible to decision-makers, we developed a "summary of summaries" approach. Given the considerable volume of information, it is impractical for managers to review each summary comprehensively. Consequently, aggregated "summary of summaries" were formulated to distill insights into concise and interpretable formats, facilitating informed decision-making, consistent with Mayande et al. (2024). These aggregated summaries encapsulate the essence of the collected data while preserving the essential nuances for follow-up strategic planning.

Comparative analyses conducted within countries (Table 8(a)) indicate significant alterations in public sentiment before and after the specific truck strike event. The mean sentiment score in China reflects a marginal decline from 0.124 to 0.105; however, this difference is not statistically significant. This slight shift underscores an increased awareness of the vulnerabilities within the SC and the necessity for resilience in procuring critical minerals. The sentiment observed post-event reveals concerns regarding the potential worsening of pre-existing issues in the SC for critical minerals, particularly those stemming from labour strikes. In contrast, the DRC demonstrates a more considerable reduction in sentiment, decreasing from 0.101 before the event to −0.032 following it, constituting a statistically significant difference. This shift is associated with prior optimistic projections concerning technological advancements, such as the development of cobalt-free batteries, which could affect both cobalt demand and pricing, thereby potentially diminishing the DRC's bargaining position. In the aftermath of the event, the strike intensified prevailing anxieties related to the DRC's mining sector, explicitly surrounding instability, thereby complicating the pursuit of ethical and sustainable cobalt mining practices. For the U.S.A. and Canada, sentiment experienced a modest decline from 0.347 to 0.314 following the event, with this difference deemed statistically significant. This decrease is attributed to an enhanced awareness regarding the dependence of the SC on transportation networks and the implications of the strike, which accentuates vulnerabilities within production and distribution processes that may provoke price surges and delays, especially in the context of EV manufacturing.

To assess answers for RQ#2 and RQ#3, we set up cross-national comparisons. These nations represent three different echelons of the cobalt SC, as shown in Figure 2

Table 8. Sentiment results and allied interpretations: (a) within-country comparisons, and (b) across-country comparisons.

(a) Within-country Comparisons					
Scenario (within country)	Before	After	Difference in sentiment mean	Difference significant	Explanations for the difference in sentiment means based on interpretation of "summary of summaries"
	Sentiment mean	Sentiment mean			
China (Before vs. After)	0.124	0.105	−0.019	No	*After event – Time lag view: <i>"There is a sense of urgency and growing awareness of the need to secure a more resilient supply chain for critical minerals"</i> .
DRC (Before vs. After)	0.101	−0.032	−0.133	Yes	*Truck strike impact mentioned (i.e., strike could exacerbate existing SC vulnerabilities). *Before event – Technological advancements perspective: <i>"The development of cobalt-free or low-cobalt batteries could significantly impact cobalt demand and prices, reducing DRC's leverage"</i> .
U.S.A. and Canada (Before vs. After)	0.347	0.314	−0.033	Yes	*After event – Truck strike mentioned, and the DRC vulnerabilities highlighted: <i>"The DRC remains vulnerable to instability and corruption, making it difficult to ensure ethical and sustainable cobalt mining practices"</i> . *After event – Truck strike mentioned, and SC vulnerabilities emphasised: <i>"The reliance on transportation networks highlights the vulnerability of global supply chains to disruptions (...) leading to price spikes and delays in EV production"</i> .
(b) Across-country Comparisons					
Scenario (across country)	Country 1	Country 2	Difference in sentiment mean	Difference significant	Explanations for the difference in sentiment means based on interpretation of "summary of summaries"
	Sentiment mean	Sentiment mean			
China vs. DRC	0.105	−0.032	−0.137	Yes	*China – Concerns about country's dominance in the field: <i>"China's dominance raises concerns about the potential for supply chain disruptions, price manipulation, and geopolitical leverage"</i> . *DRC – In general DRC own instability, also affected by the price manipulation, and geopolitical impacts.
China vs. U.S.A. and Canada	0.105	0.314	0.209	Yes	*China – Cost competitiveness perspective: <i>"China's established infrastructure and processing capabilities, along with its use of lower-cost labour and less stringent environmental regulations, make it difficult for Western companies to compete on price"</i> . *U.S.A. and Canada – Attempts to find the substitutes for cobalt in battery chemistry: <i>"Companies are exploring new battery technologies such as lithium-iron phosphate batteries and sodium-ion batteries"</i> .
DRC vs. U.S.A. and Canada	−0.032	0.314	0.346	Yes	*DRC – Concerns over responsible sourcing of cobalt in DRC. Energy transition challenges: <i>"The global demand for cobalt is expected to continue increasing as the world transitions to clean energy technologies"</i> . *U.S.A. and Canada – Recycling opportunities: <i>"Companies are developing innovative recycling technologies to recover cobalt from end-of-life batteries, minimising reliance on mining"</i> .

(with DRC focused on mining, China focused on processing and manufacturing, and the U.S.A. and Canada focused on assembly). There is clear evidence of information asymmetry across these echelons, as shown both in the results (i.e., in Table 3, with differing sample sizes) and in the "summary of summaries" analysis (i.e., in Table 8(b) data and explanation about how the information about truck strike disruption propagated). Table 8(b) offers valuable insights into each region's distinct challenges and strengths in managing its role within the global SC of critical minerals. The comparison between China and the DRC revealed a notable sentiment gap, with a post-event difference of -0.137 , which is statistically significant. While China's predominance in the cobalt SC raises concerns regarding potential disruptions and geopolitical influence, the DRC's difficulties are more closely linked to internal instability and external price manipulation, further complicating its position in the market. Moreover, the comparison between China and the U.S.A. and Canada showed a significant difference in sentiment (0.209), as China capitalises on cost advantages attributed to its established infrastructure, lower labour costs, and more lenient environmental regulations, posing challenges for Western companies to compete effectively on price. Conversely, the U.S.A. and Canada region is proactively exploring alternative battery technologies, including lithium-iron phosphate and sodium-ion batteries, to reduce reliance on cobalt. Finally, the comparison between the DRC and the U.S.A. and Canada indicated the most significant sentiment gap (0.346), highlighting the DRC's concerns regarding responsible cobalt sourcing amid global demands for ethical practices. In contrast, the U.S.A. and Canada region emphasises cobalt recycling technologies to recover cobalt from used batteries, which could mitigate reliance on mining and offer a sustainable path forward in the global energy transition. Figure 6 provides additional details on the trend of sentiments by showing the monthly average sentiment shifts by region over the 12-month time window, with and without contextualising news summaries for sentiment analysis. The analysis presented in Figure 6 is discussed in Appendix 3 for comprehensive reference and detailed examination.

6. Discussion

This section elucidates the principal findings pertinent to our approach to sentiment analysis within the domain of critical mineral SCs. First, we outline managerial decision-making opportunities stemming from our research. Subsequently, we examine the theoretical contributions of our research. Moreover, we underscore the importance of human oversight to guarantee expert

alignment with automated outcomes. We finally discuss limitations and future research avenues.

6.1. Decision opportunities

Insights from our survey of critical mineral SC stakeholders, conducted during the '2024 Cambridge Critical Mineral Supply Chain Resilience Workshop', were revelatory about the role of LLM-driven sentiment analysis in SC decision-making. The implications of employing sentiment analysis in decision-making within SCs are as follows.

First, sentiment analysis can serve as an early warning system for SC managers to identify bottlenecks and vulnerabilities. By monitoring sentiment trends, managers can foresee upstream disruptions (e.g., a DRC truck strike or geopolitical tensions) propagating to midstream and downstream operations. To a greater extent, CESCSS integrates long-context LLMs to capture the pathways through which disruptions propagate. For example, a delay in the DRC's cobalt extraction could generate negative sentiment among midstream processing facilities in China, signalling potential bottlenecks before they escalate into production delays downstream. This predictive capability resembles techniques employed in lightning equipment SCs, where ripple effect modelling can enable managers to mitigate systemic risks by identifying and preempting vulnerabilities (Ivanov 2020). Therefore, as an example, if the sentiment is positive (negative), then managers can better understand the mechanism behind forecasts that alter the demand upwards (downwards) and adjust stocking decisions, order sizes, and capacity allocation accordingly. The latter aligns with the informants' perspective in the critical mineral SC survey:

Sentiment about buying behaviours of customers could affect inventory management policies – for example, stocking decisions, order size, and capacity allocation. [Informant #17]

Secondly, businesses ought to incorporate consumer sentiments regarding environmental, social, and governance downstream SC concerns into their operational frameworks and invest in initiatives that resonate with customer preferences, such as ethical sourcing, transparent sustainability reporting, and sustainable manufacturing practices. In this spirit, an industry professional also stressed that:

News on human rights/ESG violations can reduce a company's stock price, with the market anticipating fines and additional costs (e.g., increased due diligence). [Informant #17]

Third, financial analysts and businesses could integrate real-time sentiment and market signals into risk

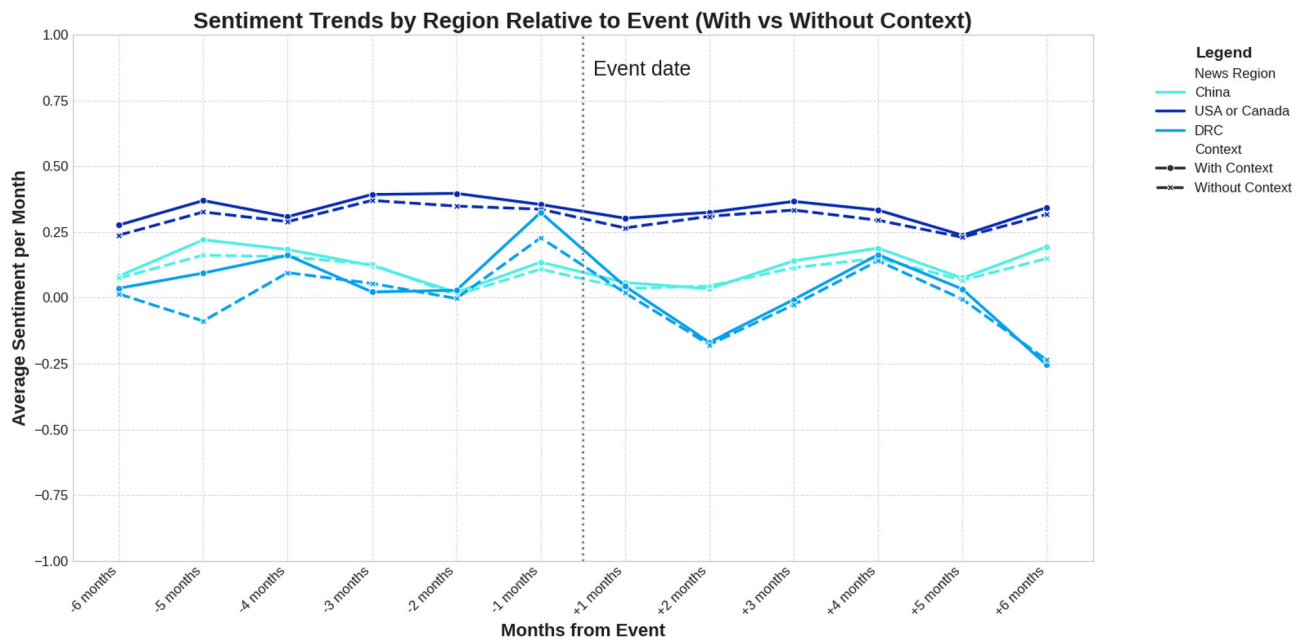


Figure 6. Monthly average sentiment shifts by region over 12 months, with and without contextualising news summaries for sentiment analysis.

assessment models to assess the potential impact of end-to-end SC disruptions on commodity valuation, e.g., shipping rates and insurance trends can act as proxies for understanding cost increases due to SC bottlenecks. One of our informants indicated that:

Potential disruptions lead to potential costs, hence impacting the valuation of the commodities via quotation of shipping rates of impacted companies and insurers, etc. [Informant #1]

In addition, financial analysts can leverage sentiment analysis to refine valuation models for critical mineral commodities and navigate market cycles more effectively, balancing immediate opportunities with long-term strategies. For instance, during pricing booms like the lithium surge, analysts could focus on short-term market opportunities while preparing for potential corrections. As one industrial professional stated:

Market valuation (particularly on critical mineral commodities) is influenced by the combined effects of commodity pricing and supply chain sentiment. See the lithium boom and subsequent crash in pricing. Currently, where commodity pricing is strong, other drivers remain largely irrelevant, including ESG issues, etc. [Informant #24]

Fourth, the proposed CESC framework opens opportunities for policy scrutiny surrounding the cobalt SC, for instance, into sustainability, ethical sourcing practices (e.g., child labour usage), pricing dynamics, and operational resilience (Zhang, Aydin, and Parker 2022). Also, the sentiment analysis of public opinion can significantly contribute to formulating policies that address

sector-specific structural challenges via devising diplomatic and economic strategies that mitigate disruptions and enhance resilience in regions that are essential to the global SC. Indicatively, an industry professional observed:

Policymakers may need to make decisions based on news that is making them look bad politically, which may not be in the best interest of the industry/supply chain resilience. [Informant #48]

Overall, the CESC framework would enable organisations to set up real-time applications and threshold policies for triggering decisions by maintaining up-to-date datasets and establishing baseline interpretations of critical mineral SC scenarios. We underscore the role of interpretability, enabled through summaries, as an essential element of decision-making based on data from complex analytical techniques. Recent research highlighted the need for data-driven explanations of disruptive events to support decisions for SC resilience (e.g., Baryannis et al. 2019; Kosasih and Brintrup 2024). In this regard, the proposed CESC framework could effectively bridge the gap between descriptive analytics and actionable decision-making, providing SC stakeholders with a pragmatic mechanism to manage risks and bolster resilience. Interpretability would allow for agile responses to unforeseen critical minerals-related risks, empowering organisations to proactively adjust their SC configuration strategies and allied economic outcomes.

In terms of societal relevance, by implementing continuous sentiment analysis and factor-specific monitoring, for instance, concerning child labour and

environmental impacts, organisations and institutions can monitor real-time fluctuations in sentiment and promptly address public and investor concerns. Therefore, decision-makers could adapt their interpretations swiftly upon the emergence of new data, facilitating expedited anomaly detection and response in harmony with the imperative of ethically and sustainably sourced critical minerals.

6.2. Pathways for theoretical contributions

From a theory development perspective, our framework and allied results delineate pathways that are analogous to the process outlined in a recent investigation for signal asymmetry on diversity, equity, and inclusion performance in SC (Li et al. 2024) based on text analyses of endogenous announcements by 161 firms. Bollen, Mao, and Zeng (2011) highlighted the predictive capability of collective sentiment when it aligns with contextual changes, thereby reinforcing the importance of context in such analytical evaluations. Li et al. (2024) introduced two differentiated constructs based on data generated (endogenously) by firms: signal strength and signal specificity. Further, the authors argued that firms should manage each construct to yield better capital market outcomes. In difference with Li et al. (2024), the source of news in our approach is exogenous, and the resulting sentiments (as indicated by the survey outlined in Section 3) may affect the mental models for SC decision-makers around how demand is predicted, risk is managed, or the capital market ascribes value. The need to explain the mechanisms behind these mental models and subsequent decisions opens theory questions. For instance, what are the mechanisms that lead to contextual factors playing a significant role (e.g., sentiment in China compared to the U.S.A. and Canada vs. the DRC) while evaluating shifts in sentiments across various SC echelons?

In Tables 3 and 4, for instance, Scenario 14 compares sentiment before the strike between China (sample size = 735) and the U.S.A. and Canada (sample size = 1,963), revealing a highly significant difference in sentiment (p -value = $9.85\text{E-}37$). This indicates that sentiment in China is significantly more negative than in the U.S.A. and Canada, reflecting the contrasting economic expectations and stability in these regions. Similarly, Scenario 15 compares sentiment before the strike between the DRC (sample size = 136) and the U.S.A. and Canada (sample size = 1,963), with a p -value of $1.42\text{E-}08$. This significant difference shows that sentiment in the DRC is notably lower, possibly due to the instability and challenges faced by the DRC's cobalt mining sector. In the post-strike period, Scenario 16 compares sentiment between China (sample size = 538) and the

DRC (sample size = 127), with a significant positive difference (p -value = 0.01), suggesting that sentiment in China remains relatively stable and positive compared to the DRC, which experiences a significant decline in sentiment due to the disruptions caused by the truck strike and ongoing instability in cobalt production. These shifts can beget the need to explain the sentiments associated with signal strength and specificity explored by Li et al. (2024), leading to the following proposition:

Proposition #1 – Contextual factors and the specificity of exogenous signals significantly influence news sentiment estimated by LLM.

In addition, precision in sentiment analysis is paramount for theory development, especially when interpreting context to inform resilience strategies within a complex SC environment affected by critical external and internal disruptions. Sentiment shifts are a dynamic aspect of communication, particularly in complex discussions like SCs, where emotional tone may change in response to evolving circumstances. Stakeholders often express fluctuating sentiments in SCs due to disruptions, uncertainties, and changing market conditions. Although the shifts in sentiment may be minimal, the asymmetry signals an inability to observe changes with precision (Wu 2023). Our statistical analyses in Tables 3 and 4 demonstrate the high accuracy (to the order of 10^{-2}) with which shifts in sentiments within corpora can be captured, illustrating significant impacts and emphasising the intricate ways context influences sentiment trends. As a result, we have the following proposition:

Proposition #2 – The precision of sentiment analysis carried out by LLM, enhanced by contextual data interpretation based on summaries, significantly improves the ability to identify and assess sentiment shifts aimed at developing effective resilience strategies in complex SC environments before, during, and after disruptions.

Moreover, understanding sentiment shifts throughout SC processes can provide valuable insights into the decision mindset within SCs and the diverse stakeholders' perspectives, ultimately helping tackle information asymmetry challenges across operations echelons. Sentiment spikes across SC echelons can inform reconfigurations to prioritise stabilising upstream operations or securing alternative supply routes. Natural language processing-based sentiment analysis was used during the COVID-19 pandemic and the Ukraine-Russia War, enabling proactive measures in SC management to efficiently plan inventories and allocate resources at the organisational capacity (Sadeek and Hanaoka 2023). The proposed CESC framework helps mitigate information discrepancies by analysing sentiment changes across SC echelons. For post hoc analysis, we asked executives

and policymakers to comment on the implications of the signalled asymmetry between China and the U.S.A. and Canada (as shown in row 2 of Table 8(b)). They suggested that these differences could significantly influence SC decision-making strategies, particularly in the areas of market valuation, demand prediction, and risk management. In the case of the U.S.A. and Canada, for example, they mentioned that "*consumer interest around supply chain decarbonisation and environmentally friendly products would impact market valuations*" [Informant #38]. If there are sentiments that new battery technologies are experiencing significant breakthroughs, "*financial markets may respond by driving prices upward*" [Informant #48]. On the other hand, China's focus on cost efficiency may stabilise valuations but make them vulnerable to regulatory changes. Executives also mentioned that this asymmetry might lead SC managers to "*start collecting data from multiple sources to be able to assess risk, compliance issues, and regulatory challenges*" [Informant #37]. To this end, we posit that:

Proposition #3 – Information asymmetry across SC echelons exacerbates LLM estimated sentiment shifts, and an increase in asymmetry is associated with an increase in the variance of the predicted sentiments.

These three propositions feature measurable constructs and are amenable to follow-on industry studies using LLMs and analytical extensions, such as identifying decision thresholds for the variety of decisions outlined in Section 6.1.

6.3. Human scrutiny for reliability and interpretability

Multidimensional reliability testing, exemplified using several statistical tests and metrics like Cronbach's alpha, requires the simultaneous evaluation of multiple factors. To validate the news article summary and news sentiment score generated by Gemini 1.5 Flash, we implemented a comprehensive two-part expert validation process. The qualitative assessment conducted by the experts sought to confirm that the summaries generated by the LLM adhered to high standards of readability and interpretability, thereby facilitating effective downstream analysis. Determining the appropriate level of human involvement in these evaluations is a critical managerial choice. While the participation of eight trained evaluators may initially appear limited, their extensive academic backgrounds and targeted training effectively mitigate interpretative variability (Landis and Koch 1977). Each evaluator conducted their assessments independently to avoid bias; however, the cognitive demands of evaluating multiple parameters and contextual layers underscore

the inherent limitations of human scrutiny. Challenges in interpretation notably arise when distinguishing sentiment shifts influenced by interdependent factors, such as cultural and regional variables. These challenges highlight the cognitive burden individuals experience in sentiment analysis, mainly when tasked with discerning meaning across thousands of texts – a complexity that Reinforcement Learning from Human Feedback may assist with.

6.4. Limitations

Our approach does not fix the hallucination problem in a foundational LLM model. Instead, we put in a scrutiny mechanism and facilitate 'multi-shot' learning within the prompt engineering process (Chen et al. 2023). However, given the sampling approach employed for interpretation, there remains a possibility that systematic bias could be introduced, potentially eluding detection through the human-centred elements of our process. The second notable limitation of our approach in applying LLMs for summarisation and sentiment analysis is the potential loss of salient information within extensive context windows. As underscored by Liu et al. (2023b), LLMs frequently encounter difficulties in effectively retrieving critical information embedded within the central sections of lengthy text passages. Contemporary LLMs often prioritise information at the introduction or conclusion of a text, which may lead to the accidental omission of vital data elsewhere within the main corpora body.

Furthermore, a technical report from Google regarding the performance of the Gemini 1.5 Pro model further elucidates this limitation. Specifically, when using shorter context lengths, Gemini 1.5 Pro performs better in the 'Needle in a Haystack' test to evaluate the model's recall of specific sentences (Gemini Team 2024). As the context length extends towards one million tokens, recall rates diminish to approximately 60%, signifying that nearly 40% of pertinent information may be 'lost' as the model's ability to retain context diminishes.

An additional limitation emerges from the variations in data sources, particularly the disparities between datasets in English and those in other languages. While this study predominantly utilises a dataset comprising approximately 5,168 English-language texts, a comparable collection from Hamborg, Meuschke, and Gipp (2020) consists of another 10,000 articles in various languages. Relying predominantly on English-language datasets engenders a regional bias, as most news sources, such as CNN, Reuters, and BBC, are predicated in Western nations rather than on local media establishments in regions such as the DRC or China, where

alternative languages are utilised. This circumstance may distort interpretations of perspectives and trends prevalent within specific foreign cultural, political, or geographic contexts as opposed to those of the local populations. Addressing these limitations is crucial for improving the approach's performance in information retrieval within extensive context windows and ensuring balanced cultural representation across diverse linguistic datasets.

6.5. Future research

Further research should also investigate the scalability and adaptability of LLM-based sentiment analysis across multilingual corpora. Since news articles and stakeholders' communication frequently exist in multiple languages, integrating translation models with sentiment analysis could enhance the robustness of insights in a global SC setting. Another interesting research avenue is mitigating hallucinations and biases in AI-driven sentiment assessments. While this study incorporates human validation mechanisms, it could benefit from more sophisticated techniques. These include RLHF and hybrid AI-human decision-support systems, which may further improve interpretability and reliability. Finally, exploring real-time sentiment monitoring for predictive decision-making offers an exciting opportunity. Implementing a dynamic, continuously updated framework could enable SC managers to respond proactively to disruptions by integrating live data streams from news outlets, financial markets, and social media. For example, our CESCSS framework could be applied to analyse the impact of the ongoing stream of announcements regarding tariff impositions on exchange rates (Matveev and Ruge-Murcia 2024) and global sourcing (Zhang, Dabadghao, and Udenio 2025) in critical resource SCs. For instance, our CESCSS framework provides feedback in aggregate, as shown in Figure 1. It is possible to provide feedback based on either factor analysis or segments (e.g., by country). This represents a potential improvement for the CESCSS framework, which would facilitate its adaptation to specific factors or countries. Investigating how real-time insights affect SC resilience strategies and market behaviours would enhance the practical applicability of AI-driven sentiment analysis in Operations Management.

7. Conclusion

The absence of information sharing among multi-echelon SC stakeholders reduces efficiency due to information deficiencies, which affects trust (Inderfurth, Sadrieh, and Voigt 2013). Utilising LLMs and summaries

for sentiment analysis addresses this issue of information asymmetry and deficits in the following ways. Data and inferences from LLMs are being developed and are being scrutinised in different academic disciplines. We presented an exploratory study to employ an LLM to assess SC sentiments and to interpret shifts associated with such sentiments in the event of disruption. Current approaches to SC sentiment analysis predominantly rely on lexicon-based or traditional machine learning approaches, limiting their ability to capture contextual nuances effectively. An LLM with a long context window offers the opportunity to incorporate contextual information into sentiment analysis. However, we caution against treating LLMs as 'black boxes'. Context, interpretation, and human feedback are essential for accurate sentiment assessment and for buy-in from stakeholders. The CESCSS framework effectively extracts context and sentiment utilising LLMs, potentially aiding decision-making for improving SC resilience (Cohen et al. 2022; Craighead et al. 2007). Our work can complement existing analyses of cost-effective countermeasures to SC disruptions (Simchi-Levi, Trichakis, and Zhang 2019). We see the CESCSS framework as a first step toward integrating extensive text data into meaningful SC observations, tackling hallucination challenges while acknowledging the need to address its limitations and pursue further theoretical investigations.

Acknowledgements

An early version of this manuscript was reviewed by David Anderson (Villanova University), Sharookh Daruwalla (Tellagence), Nitin Mayade (Tellagence), and Sriram Narayan (Michigan State University). We are grateful for their insightful feedback. Sharookh Daruwalla and Nitin Mayade also provided valuable input on best practices and potential pitfalls in the use of Large Language Models (LLMs). We would also like to acknowledge the contributions of Fangrui Li, Maharshi Dhada, Thorge Schaum, Garry Clawson, Yongxu Zhao, Zhong Guan, and Abhishek Agarwal from the University of Cambridge, as well as Wan Ri Ho from ETH Zurich, for their support. Dr. Naoum Tsolakis gratefully acknowledges his affiliation with the Postdoctoral Society of Trinity College, University of Cambridge, whose stimulating intellectual environment contributed to the development of this research.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research is part of the UK-India Critical Mineral Supply Chain Observatory project at the University of Cambridge, funded by the UK's Department for Science, Innovation and Technology (DSIT), Institution Reference: G126878.

Notes on contributors



Nitin Joglekar holds the John F. Connelly Endowed Chair in Management at Villanova School of Business. His research interests span (i) Human-centered AI aimed at platforms, products, and supply chain innovations, and (ii) Agentic-AI coordination, governance and transparency risks in Supply chains.



Wei Nie is a PhD researcher at the Industrial Resilience Research Group and the Lead Researcher in Big Data Analytics for the Global Supply Chain Observatory (GSCO) at the Institute for Manufacturing, Department of Engineering, University of Cambridge. Her research focuses on leveraging the publicly available data to develop the end-to-end supply chain visibility, with a particular emphasis on critical mineral supply chains. Wei has professional experience as a data scientist and software engineer in the manufacturing sector, with a focus on the semiconductor industry. She is interested in data and systems modelling & simulation. Wei holds an MPhil in Industrial Systems, Manufacture, and Management from the University of Cambridge and a BEng (Honours) in Industrial & Systems Engineering from the National University of Singapore.



Mariel Alem Fonseca is a PhD researcher at the Industrial Resilience Research Group and Lead Researcher in Supply Chain Resilience for the Global Supply Chain Observatory (GSCO) at the Institute for Manufacturing, Department of Engineering, University of Cambridge. Her research focuses on the trade-offs between supply chain resilience and cost, with a particular focus on emerging food industries such as alternative proteins. She has over 6 years of experience in process design and optimisation, supply chain management and data analytics. Mariel holds a Master in Engineering for Sustainable Development from the University of Cambridge and an undergraduate degree in Industrial Engineering from the University of Costa Rica.



Naoum Tsolakis is an Associate Professor of Supply Chain and Operations Management at the International Hellenic University (Greece) and a Research Collaborator at the Institute for Manufacturing, University of Cambridge (UK). His research focuses on modelling and analysing supply chain resilience, integrating System Dynamics, agent-based simulation, and AI-driven methods to support informed decision-making in the face of disruptions. He holds a Ph.D. in Mechanical Engineering from the Aristotle University of Thessaloniki and Master's degrees from the University of Thessaly, University College London, and the Hellenic Open University, in areas including engineering design methods for industrial systems, technology entrepreneurship, project management, and environmental infrastructure. He has published extensively in top-tier journals and contributes to EU-funded research on sustainable and digitally enabled supply networks.



Mukesh Kumar received the Ph.D. degree in industrial engineering from University of Cambridge, in 2010. He is an Associate Professor in Operations Management, Head of the Industrial Resilience Research Group and the Director of Global Supply Chain Observatory (GSCO) at the University of Cambridge. His academic tenure at Cambridge has been marked by a focus on resilience, sustainability, and global supply chains. Prior to his academic career, he gained substantial experience in corporate finance and investment banking in Mumbai and Bangkok. He teaches courses ranging from supply chain risk management to industrial psychology. His research interests include the U.K.–India critical mineral supply chain observatory and circular economy practices in manufacturing networks, all of which have attracted substantial grant support. Dr. Kumar was the recipient of the Highly Commended Paper at EUROMA and the Best Paper Award at the International Journal of Production Research.

Data availability statement

Data available within the article or its supplementary materials.

ORCID

Nitin Joglekar  <http://orcid.org/0000-0002-6478-337X>

Wei Nie  <http://orcid.org/0000-0002-1419-2125>

Mariel Alem Fonseca  <http://orcid.org/0000-0003-2758-3608>

Naoum Tsolakis  <http://orcid.org/0000-0003-2042-7047>

Mukesh Kumar  <http://orcid.org/0000-0002-1961-5078>

References

- Ajwani, R., S. R. Javaji, F. Rudzicz, and Z. Zhu. 2024. "LLM-Generated Black-Box Explanations can be Adversarially Helpful." <https://doi.org/10.48550/arXiv.2405.06800>.
- Anderson, D., S. Mahar, and S. Modi. 2024. *ChatGPT Reads Supply Chain News: Using Large Language Models to Forecast Price Index Changes*. Working Paper, Villanova University.
- Bakker, M. A., M. J. Chadwick, H. R. Sheahan, M. H. Tessler, J. Balaguer, N. McAleese, A. Glaese, J. Aslanides, M. M. Botvinick, and C. Summerfield. 2022. "Fine-Tuning Language Models to Find Agreement Among Humans with Diverse Preferences." <https://doi.org/10.48550/arXiv.2211.15006>.
- Ball, G. P., H. Bavafa, C. C. Blanco, H. Park, and K. D. Wowak. 2024. "Gender and Serious Drug Recalls: A Textual Sentiment Analysis of Drug Reviews on WebMD." *Production and Operations Management* 34 (4): 698–710. <https://doi.org/10.1177/10591478241256644>.
- Baryannis, G., S. Validi, S. Dani, and G. Antoniou. 2019. "Supply Chain Risk Management and Artificial Intelligence: State of the Art and Future Research Directions." *International Journal of Production Research* 57 (7): 2179–2202. <https://doi.org/10.1080/00207543.2018.1530476>.
- Bau, R. T. R. L., A. Hermila, and M. Latief. 2024. "Electric Vehicle Adoption in Indonesia: Insights from Sentiment Analysis of Major News Portals." In *2024 International Conference on Data Science and Its Applications (ICoDSA)*. IEEE.

- Belhadi, A., S. Kamble, S. Fosso Wamba, and M. M. Queiroz. 2021. "Building Supply-Chain Resilience: An Artificial Intelligence-Based Technique and Decision-Making Framework." *International Journal of Production Research* 60 (14): 4487–4507. <https://doi.org/10.1080/00207543.2021.1950935>.
- Beyer Díaz, S., K. Coussement, A. De Caigny, L. F. Pérez, and S. Creemers. 2024. "Do the US President's Tweets Better Predict Oil Prices? An Empirical Examination Using Long Short-Term Memory Networks." *International Journal of Production Research* 62 (6): 2158–2175. <https://doi.org/10.1080/00207543.2023.2217286>.
- Bollen, J., H. Mao, and X. Zeng. 2011. "Twitter Mood Predicts the Stock Market." *Journal of Computational Science* 2 (1): 1–8. <https://doi.org/10.1016/j.jocs.2010.12.007>.
- Chae, B. 2015. "Insights from Hashtag #Supplychain and Twitter Analytics: Considering Twitter and Twitter Data for Supply Chain Practice and Research." *International Journal of Production Economics* 165:247–259. <https://doi.org/10.1016/j.ijpe.2014.12.037>.
- Chen, B., D. Simchi-Levi, Y. Wang, and Y. Zhou. 2021. "Dynamic Pricing and Inventory Control with Fixed Ordering Cost and Incomplete Demand Information." *Management Science* 68 (8): 5684–5703. <https://doi.org/10.1287/mnsc.2021.4171>.
- Chen, B., Z. Zhang, N. Langrené, and S. Zhu. 2023. "Unleashing the Potential of Prompt Engineering in Large Language Models: A Comprehensive Review." <https://doi.org/10.48550/arXiv.2310.14735>.
- Chohan, A. 2016. *Data in Government: Analysing Survey Comments Using Sentiment Scores*. Accessed 29 October 2024. <https://dataingovernment.blog.gov.uk/2016/02/18/analysing-survey-comments-using-sentiment-scores/>.
- Cobalt Institute. 2024. *Cobalt Market Report 2023*. London. Accessed 29 October 2024. <https://nora.nerc.ac.uk/id/eprint/537241/>.
- Cohen, M., S. Cui, S. Doetsch, R. Ernst, A. Huchzermeier, P. Kouvelis, H. Lee, H. Matsuo, and A. A. Tsay. 2022. "Bespoke Supply-Chain Resilience: The Gap Between Theory and Practice." *Journal of Operations Management* 68 (5): 515–531. <https://doi.org/10.1002/joom.1184>.
- Craighead, C. W., J. Blackhurst, M. J. Rungtusanatham, and R. B. Handfield. 2007. "The Severity of Supply Chain Disruptions: Design Characteristics and Mitigation Capabilities." *Decision Sciences* 38 (1): 131–156. <https://doi.org/10.1111/j.1540-5915.2007.00151.x>.
- Cuberek, R., M. Janíková, and J. Dygrýn. 2021. "Adaptation and Validation of the Physical Activity Questionnaire for Older Children (PAQ-C) Among Czech Children." *PLoS One* 16 (1): e0245256. <https://doi.org/10.1371/journal.pone.0245256>.
- Devlin, J., M. Chang, K. Lee, and K. Toutanova. 2018. *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. <https://doi.org/10.48550/arXiv.1810.04805>.
- Dey, P. K., S. Chowdhury, A. Abadie, E. Vann Yarosan, and S. Sarkar. 2023. "Artificial Intelligence-Driven Supply Chain Resilience in Vietnamese Manufacturing Small- and Medium-Sized Enterprises." *International Journal of Production Research* 62 (15): 5417–5456. <https://doi.org/10.1080/00207543.2023.2179859>.
- Dow Jones. 2024. *Factiva – Global News Monitoring, Business Intelligence Platform*. Accessed 22 January 2025. <https://www.dowjones.com/professional/factiva/>.
- Faruquee, M., A. Paulraj, and C. A. Irawan. 2023. "A Typology of Supply Chain Resilience: Recognising the Multi-capability Nature of Proactive and Reactive Contexts." *Production Planning and Control* 35 (12): 1503–1523. <https://doi.org/10.1080/09537287.2023.2202151>.
- Gaddala, V. S. 2023. "Unleashing the Power of Generative AI and RAG Agents in Supply Chain Management: A Futuristic Perspective." *IRE Journals* 6 (12): 1411–1417.
- Gadelho, C. 2024. *RAG vs Large Context Models: How Gemini 1.5 Changes the world*. TensorOps. Accessed 20 October 2024. <https://www.tensorops.ai/post/rag-vs-large-context-models-how-gemini-1-5-changes-the-world>.
- Gemini Team. 2024. *Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context*. Google. <https://doi.org/10.48550/arXiv.2403.05530>.
- Geng, R., H. K. S. Lam, and M. Stevenson. 2022. "Addressing Modern Slavery in Supply Chains: An Awareness-Motivation-Capability Perspective." *International Journal of Operations & Production Management* 42 (3): 331–356. <https://doi.org/10.1108/IJOPM-07-2021-0425>.
- Golicic, S. L., and D. F. Davis. 2012. "Implementing Mixed Methods Research in Supply Chain Management." *International Journal of Physical Distribution and Logistics Management* 42 (8/9): 726–741. <https://doi.org/10.1108/09600031211269721>.
- Google AI for Developers. 2024. *Long Context*. Google. Accessed 20 October 2024. <https://ai.google.dev/gemini-api/docs/long-context>.
- Gupta, V., P. N. Venkit, S. Wilson, and R. J. Passonneau. 2024. *Sociodemographic Bias in Language Models: A Survey and Forward Path*. <https://doi.org/10.48550/arXiv.2306.08158>.
- Hamborg, F., N. Meuschke, and B. Gipp. 2020. "Bias-aware News Analysis Using Matrix-Based News Aggregation." *International Journal on Digital Libraries* 21:129–147. <https://doi.org/10.1007/s00799-018-0239-9>.
- Hazen, B. T., C. A. Boone, J. D. Ezell, and L. A. Jones-Farmer. 2014. "Data Quality for Data Science, Predictive Analytics, and Big Data in Supply Chain Management: An Introduction to the Problem and Suggestions for Research and Applications." *International Journal of Production Economics* 154:72–80. <https://doi.org/10.1016/j.ijpe.2014.04.018>.
- Huang, L., W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, et al. 2025. "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions." *ACM Transactions on Information Systems* 43 (2): 1–55. <https://doi.org/10.1145/3703155>.
- Inderfurth, K., A. Sadrieh, and G. Voigt. 2013. "The Impact of Information Sharing on Supply Chain Performance Under Asymmetric Information." *Production and Operations Management* 22 (2): 410–425. <https://doi.org/10.1111/j.1937-5956.2012.01372.x>.
- Ivanov, D. 2020. "Predicting the Impact of Epidemic Outbreaks on Global Supply Chains: A Simulation-Based Analysis on the Coronavirus Outbreak (COVID-19/SARS-CoV-2) Case." *Transportation Research Part E: Logistics and Transportation Review* 136:101922. <https://doi.org/10.1016/j.tre.2020.101922>.

- Ivanov, D. 2022. "Blackout and Supply Chains: Cross-Structural Ripple Effect, Performance, Resilience and Viability Impact Analysis." *Annals of Operations Research* 348: 1127–1143. <https://doi.org/10.1007/s10479-022-04754-9>.
- Jackson, I., D. Ivanov, A. Dolgui, and J. Namdar. 2024. "Generative Artificial Intelligence in Supply Chain and Operations Management: A Capability-Based Framework for Analysis and Implementation." *International Journal of Production Research* 62 (17): 6120–6145. <https://doi.org/10.1080/00207543.2024.2309309>.
- Jackson, I., M. Jesus Saenz, and D. Ivanov. 2023. "From Natural Language to Simulations: Applying AI to Automate Simulation Modelling of Logistics Systems." *International Journal of Production Research* 62 (4): 1434–1457. <https://doi.org/10.1080/00207543.2023.2276811>.
- Karl, A., J. Wisnowski, and W. H. Rushing. 2015. "A Practical Guide to Text Mining with Topic Extraction." *Wiley Interdisciplinary Reviews: Computational Statistics* 7 (5): 326–340. <https://doi.org/10.1002/wics.1361>.
- Kleindorfer, P. R., and G. H. Saad. 2005. "Managing Disruption Risks in Supply Chains." *Production and Operations Management* 14 (1): 53–68. <https://doi.org/10.1111/j.1937-5956.2005.tb00009.x>.
- Kosasih, E. E., and A. Brintrup. 2022. "A Machine Learning Approach for Predicting Hidden Links in Supply Chain with Graph Neural Networks." *International Journal of Production Research* 60 (17): 5380–5393. <https://doi.org/10.1080/00207543.2021.1956697>.
- Kosasih, E. E., and A. Brintrup. 2024. "Towards Trustworthy AI for Link Prediction in Supply Chain Knowledge Graph: A Neurosymbolic Reasoning Approach." *International Journal of Production Research* 63 (6): 2268–2290. <https://doi.org/10.1080/00207543.2024.2399713>.
- Kosasih, E. E., E. Papadakis, G. Baryannis, and A. Brintrup. 2024. "A Review of Explainable Artificial Intelligence in Supply Chain Management Using Neurosymbolic Approaches." *International Journal of Production Research* 62 (4): 1510–1540. <https://doi.org/10.1080/00207543.2023.2281663>.
- Kumar, V. V., A. Sahoo, S. K. Balasubramanian, and S. Gholston. 2025. "Mitigating Healthcare Supply Chain Challenges Under Disaster Conditions: A Holistic AI-Based Analysis of Social Media Data." *International Journal of Production Research* 63 (2): 779–797. <https://doi.org/10.1080/00207543.2024.2316884>.
- Lahat, A., E. Shachar, B. Avidan, Z. Shatz, B. S. Glicksberg, and E. Klang. 2023. "Evaluating the Use of Large Language Model in Identifying Top Research Questions in Gastroenterology." *Scientific Reports* 13 (1): 1–6. <https://doi.org/10.1038/s41598-023-31412-2>.
- Landis, J. R., and G. G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33 (1): 159–174. <https://doi.org/10.2307/2529310>.
- Li, F., C. K. Lo, C. S. Tang, and P. Zhou. 2024. "Will Diversity, Equity, and Inclusion Commitment Improve Manufacturing Firms' Market Performance? A Signaling Theory Perspective on DEI Announcements." *Production and Operations Management* 34 (3): 331–342. <https://doi.org/10.1177/10591478241265479>.
- Li, H., Y. Su, D. Cai, Y. Wang, and L. Liu. 2022. "A Survey on Retrieval-Augmented Text Generation." <https://doi.org/10.48550/arXiv.2202.01110>.
- Li, M., X. Zhang, and B. Dan. 2019. "Competition and Cooperation in a Supply Chain with an Offline Showroom Under Asymmetric Information." *International Journal of Production Research* 58 (19): 5964–5979. <https://doi.org/10.1080/00207543.2019.1661536>.
- Liu, Y., D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. 2023a. "G-Eval: NLG Evaluation Using GPT-4 with Better Human Alignment." <https://doi.org/10.48550/arXiv.2303.16634>.
- Liu, N. F., K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. 2023b. "Lost in the Middle: How Language Models Use Long Contexts." <https://doi.org/10.48550/arXiv.2307.03172>.
- Liu, Y., Y. Yao, J. F. Ton, X. Zhang, R. G. H. Cheng, Y. Klochkov, M. F. Taufiq, and H. Li. 2023c. "Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment." <https://doi.org/10.48550/arXiv.2308.05374>.
- Loughran, T., and B. McDonald. 2011. "When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *The Journal of Finance* 66:35–65. <https://doi.org/10.1111/j.1540-6261.2010.01625.x>.
- Luk, J., and F. Njini. 2023. "Copper, Cobalt Supplies from Congo Stranded by Truckers' Strike." *Reuters*. Accessed 12 September 2024. <https://www.reuters.com/markets/commodities/copper-cobalt-supplies-congo-stranded-by-truckers-strike-2023-11-07/#:~:text=%22It's%20affecting%20most%20of%20the,value%20copper%20and%20cobalt%20materials>.
- Matveev, D., and F. Ruge-Murcia. 2024. "Tariffs and the Exchange Rate: Evidence from Twitter." *IMF Economic Review* 72:1185–1211. <https://doi.org/10.1057/s41308-023-00206-x>.
- Mayande, N., S. Daruwalla, S. Khodke, N. Joglekar, and C. Weber. 2024. "Syntactic and Semantic Attention Summary (SSAS): An Approach to Improve LLM Summary Generation." In *INFORMS 2024 Annual Meeting*, Seattle, WA.
- McCloskey, B. J., P. M. LaCasse, and B. A. Cox. 2024. "Natural Language Processing Analysis of Online Reviews for Small Business: Extracting Insight from Small Corpora." *Annals of Operations Research* 341:295–312. <https://doi.org/10.1007/s10479-023-05816-2>.
- Medhat, W., A. Hassan, and H. Korashy. 2014. "Sentiment Analysis Algorithms and Applications: A Survey." *Ain Shams Engineering Journal* 5 (4): 1093–1113. <https://doi.org/10.1016/j.asej.2014.04.011>.
- Menache, I., J. Pathuri, D. Simchi-Levi, and T. Linton. 2025. "How Generative AI Improves Supply Chain Management." *Harvard Business Review* 104 (1–2): 86–95.
- Müller, R., and P. Büttner. 1994. "A Critical Discussion of Intraclass Correlation Coefficients." *Statistics in Medicine* 13 (23–24): 2465–2476. <https://doi.org/10.1002/sim.4780132310>.
- Nguyen, A., R. Pellerin, S. Lamouri, and B. Leksens. 2022. "Managing Demand Volatility of Pharmaceutical Products in Times of Disruption Through News Sentiment Analysis." *International Journal of Production Research* 61 (9): 2829–2840. <https://doi.org/10.1080/00207543.2022.2070044>.
- Özer, Ö., Y. Zheng, and K.-Y. Chen. 2011. "Trust in Forecast Information Sharing." *Management Science* 57 (6): 1111–1137. <https://doi.org/10.1287/mnsc.1110.1334>.
- Ozyurt, B., and M. A. Akcayol. 2021. "A New Topic Modeling Based Approach for Aspect Extraction in Aspect Based Sentiment Analysis: SS-LDA." *Expert Systems with*

- Applications* 168:114231. <https://doi.org/10.1016/j.eswa.2020.114231>.
- Pelon, R., K. L. Hund, T.P. Fabregas Masllovet, P. De Sa, and G.J.R. McMahon. 2021. *Cobalt in the Democratic Republic of Congo: Market Analysis (English)*. Washington, D.C: World Bank Group.
- Pratelli, M., J. Bianchi, F. Pinelli, and M. Petrocchi. 2024. "Evaluation of Reliability Criteria for News Publishers with Large Language Models." <https://doi.org/10.48550/arXiv.2412.15896>.
- Quiggin, D., and R. King. 2023. *Cobalt Refining Power Gives China an Advantage in the Race for EV Battery Dominance*. Chatham House. Accessed 9 September 2024. <https://resourcetrade.earth/publications/critical-metals-ev-batteries#:~:text=In%202022%2C%20Chinese%20companies%20accounted,10%20percentage%20points%20for%20lithium>.
- Ram, O., Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, and Y. Shoham. 2023. "In-Context Retrieval-Augmented Language Models." <https://doi.org/10.48550/arXiv.2302.00083>.
- Rudin, C. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1 (5): 206–215. <https://doi.org/10.1038/s42256-019-0048-x>.
- Sadeek, S. N., and S. Hanaoka. 2023. "Assessment of Text-Generated Supply Chain Risks considering News and Social Media During Disruptive Events." *Social Network Analysis and Mining* 13 (1): 1–18. <https://doi.org/10.1007/s13278-023-01100-0>.
- Sallam, M., N. A. Salim, M. Barakat, K. Al-Mahzoum, B. Al-Tammemi, D. Malaeb, R. Hallit, and S. Hallit. 2023. "Assessing Health Students' Attitudes and Usage of ChatGPT in Jordan: Validation Study." *JMIR Medical Education* 9: 1–15. <https://doi.org/10.2196/48254>.
- Schmidt, C. G., D. A. Wuttke, G. P. Ball, and H. S. Heese. 2020. "Does Social Media Elevate Supply Chain Importance? An Empirical Examination of Supply Chain Glitches, Twitter Reactions, and Stock Market Returns." *Journal of Operations Management* 66 (6): 646–669. <https://doi.org/10.1002/joom.1087>.
- Serrano-Guerrero, J., J. A. Olivas, F. P. Romero, and E. Herrera-Viedma. 2015. "Sentiment Analysis: A Review and Comparative Analysis of Web Services." *Information Sciences* 311:18–38. <https://doi.org/10.1016/j.ins.2015.03.040>.
- Shamshiri, A., K. R. Ryu, and J. Y. Park. 2024. "In-Context Learning for Long-Context Sentiment Analysis on Infrastructure Project Opinions." <https://doi.org/10.48550/arXiv.2410.11265>.
- Simchi-Levi, D., N. Trichakis, and P. Y. Zhang. 2019. "Designing Response Supply Chain Against Bioattacks." *Operations Research* 67 (5): 1246–1268. <https://doi.org/10.1287/opre.2019.1862>.
- Singh, N. K., S. Agal, T. R. Gadekallu, M. Shabaz, I. Keshta, L. Jindal, M. Soni, H. Byeon, and P. P. Singh. 2024a. "Deep Learning Model for Interpretability and Explainability of Aspect-Level Sentiment Analysis Based on Social Media." *IEEE Transactions on Computational Social Systems* 12 (3): 1307–1318. <https://doi.org/10.1109/TCSS.2023.3347664>.
- Singh, C., J. P. Inala, M. Galley, R. Caruana, and J. Gao. 2024b. "Rethinking Interpretability in the Era of Large Language Models." <https://doi.org/10.48550/arXiv.2402.01761>.
- Soleymani, M., D. Garcia, B. Jou, B. Schuller, S. Chang, and M. Pantic. 2017. "A Survey of Multimodal Sentiment Analysis." *Image and Vision Computing* 65:3–14. <https://doi.org/10.1016/j.imavis.2017.08.003>.
- Son, B.-G., S. Roscoe, and M. S. Sodhi. 2025. "Dynamic Capabilities of Global and Local Humanitarian Organizations with Emergency Response and Long-Term Development Missions." *International Journal of Operations & Production Management* 45 (1): 1–32. <https://doi.org/10.1108/IJOPM-12-2022-0778>.
- Sonnier, G. P., L. McAlister, and O. J. Rutz. 2011. "A Dynamic Model of the Effect of Online Communications on Firm Sales." *Marketing Science* 30 (4): 702–716. <https://doi.org/10.1287/mksc.1110.0642>.
- Sun, Z., S. Shen, S. Cao, H. Liu, C. Li, Y. Shen, C. Gan, et al. 2023. "Aligning Large Multimodal Models with Factually Augmented RLHF." <https://doi.org/10.48550/arXiv.2309.14525>.
- Taber, K. S. 2018. "The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education." *Research in Science Education* 48:1273–1296. <https://doi.org/10.1007/s11165-016-9602-2>.
- Tang, Z., D. D. Wu, and A. Dolgui. 2020. "Option Contracts for Online Celebrities as Retailers in Supply Chains." *International Journal of Production Research* 58 (14): 4215–4232. <https://doi.org/10.1080/00207543.2019.1650977>.
- Tiwari, M. K., B. Bidanda, J. Geunes, K. Fernandes, and A. Dolgui. 2024. "Supply Chain Digitisation and Management." *International Journal of Production Research* 62 (8): 2918–2926. <https://doi.org/10.1080/00207543.2024.2316476>.
- Trkman, P., K. McCormack, M. de Oliveira, and M. Ladeira. 2010. "The Impact of Business Analytics on Supply Chain Performance." *Decision Support Systems* 49:318–327. <https://doi.org/10.1016/j.dss.2010.03.007>.
- U.S. Geological Survey. 2024. *Mineral Commodity Summaries: Cobalt (January 2024)*. Accessed 29 September 2024. <https://pubs.usgs.gov/periodicals/mcs2024/mcs2024-cobalt.pdf>.
- Varathan, K. D., A. Giachanou, and F. Crestani. 2017. "Comparative Opinion Mining: A Review." *Journal of the Association for Information Science and Technology* 68 (4): 811–829. <https://doi.org/10.1002/asi.23716>.
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. "Attention is All You Need." <https://doi.org/10.48550/arXiv.1706.03762>.
- Venkit, P. N., T. Chakravorti, V. Gupta, H. Biggs, M. Srinath, K. Goswami, S. Rajtmajer, and S. Wilson. 2024. "An Audit on the Perspectives and Challenges of Hallucinations in NLP." <https://doi.org/10.48550/arXiv.2404.07461>.
- Vishnuthilak, K. K., B. Rolf, T. Reggelin, and S. Lang. 2024. "Using Sentiment Analysis to Detect Disruptive Events in Supply Chains." *IFAC-PapersOnLine* 58 (19): 864–869. <https://doi.org/10.1016/j.ifacol.2024.09.178>.
- Wolfe, R., and T. Mitra. 2024. "The Impact and Opportunities of Generative AI in Fact-Checking." <https://doi.org/10.48550/arXiv.2405.15985>.
- Wu, D. 2023. "Text-Based Measure of Supply Chain Risk Exposure." *Management Science* 70 (7): 4781–4801. <https://doi.org/10.1287/mnsc.2023.4927>.

- Xiao, Y., and W. Y. Wang. 2021. "On Hallucination and Predictive Uncertainty in Conditional Language Generation." <https://doi.org/10.48550/arXiv.2103.15025>.
- Yang, S., L. Cui, J. Xie, and Y. Zhang. 2019. "Making the Best Use of Review Summary for Sentiment Analysis." <https://doi.org/10.48550/arXiv.1911.02711>.
- Yilmaz, F. G. K., R. Yilmaz, and M. Ceylan. 2023. "Generative Artificial Intelligence Acceptance Scale: A Validity and Reliability Study." *International Journal of Human – Computer Interaction* 40 (24): 8703–8715. <https://doi.org/10.1080/10447318.2023.2288730>.
- Zhang, H., G. Aydin, and R. P. Parker. 2022. "Social Responsibility Auditing in Supply Chain Networks." *Management Science* 68 (2): 1058–1077. <https://doi.org/10.1287/mnsc.2020.3950>.
- Zhang, D. J., S. S. Dabadghao, and M. Udenio. 2025. "Global Sourcing Under Tariffs: A Perspective of Time Series Analysis." *International Journal of Production Economics* 280:109473. <https://doi.org/10.1016/j.ijpe.2024.109473>.
- Zhang, W., Y. Deng, B. Liu, S. J. Pan, and L. Bing. 2023. "Sentiment Analysis in the Era of Large Language Models: A Reality Check." <https://doi.org/10.48550/arXiv.2305.15005>.

Appendices

Appendix 1. Research context

Table A1. Prompt template for generating context.

Prompt template for generating context (summary) from individual news prior to the truck strike event with Gemini 1.5 Flash
Write a summary of this news article in the context of cobalt and supply chains: 'article'.
Prompt template for generating context (summary) from individual news after the truck strike event with Gemini 1.5 Flash
Write a summary of this news article in the context of cobalt and supply chains and/or truck strike: 'article'.

Appendix 2. Sentiment analysis process

Article summary use approach

The article summary is integral to the CESCSS framework proposed in this study. Summaries serve as a preprocessing mechanism to provide contextual information before the LLM generates sentiment data (refer to Step 5 in Figure 1). Instead of feeding full articles directly into the LLM for sentiment evaluation, summaries were generated first to capture the key contextual elements of the text and are then used as inputs for sentiment analysis, ensuring that the LLM's sentiment scoring is based on relevant information rather than less critical content within lengthy articles.

Table 2 presents the prompt template used in Gemini 1.5 Flash for sentiment analysis, utilising the article summary as context. For instance, the sentiment of a pre-event article on cobalt oversupply and price decline is generated using the following prompt: 'Given the context: 'The cobalt market is facing a significant supply glut due to increased production from Indonesia and Africa (...)', analyse the sentiment of the following texts: 'article'. Provide a sentiment score ranging from -1 to 1 (inclusive). Please only print (...)'.

This approach is supported by the Syntactic and Semantic Attention Summary (SSAS) methodology (Mayande et al. 2024), which emphasises using summaries to structure and guide LLM-based sentiment analysis. Additionally, the summary process facilitates human intervention, allowing experts to review and refine the context before sentiment assessment. This reduces potential hallucinations and improves the reliability of sentiment results.

Sentiment score validation process

The sentiment scores were validated using a structured approach that integrated expert assessments and statistical validation to ensure reliability and interpretability. The validation process comprised three key steps.

First, we assessed the reliability of sentiment scores through human expert evaluations. Eight trained SC experts were provided with a sample of six news articles and their corresponding LLM-generated sentiment scores. The sample included all three regions and represented negative, neutral, and positive sentiments to ensure the LLM's consistency across different article types. The experts independently assessed sentiment based on six key factors relevant to the cobalt SC, which were derived from factor analysis (or, more precisely, topic extraction), notably Principal Component Analysis and Latent Dirichlet Allocation. Their sentiment classifications (negative, neutral, or positive) were then converted into numerical scores for quantitative comparison with LLM-generated outputs.

Second, statistical validation was performed to compare the sentiment scores generated by human experts and those produced by the LLM. This comparison involved calculating the intraclass and Pearson correlation coefficients. Furthermore, a mixed-effects linear regression model was employed to evaluate the influence of the rater type (human vs. LLM) on sentiment scores, ensuring that any systematic biases could be identified.

Third, inter-rater reliability among human evaluators was examined using Cronbach's alpha to ensure consistency across human ratings and strengthen the reliability of the findings. In general, by integrating human assessments with statistical validation, the sentiment scores generated by the CESCSS framework were found to be both reliable and interpretable.

Appendix 3. Sentiment shifts' timeline

Figure 6 depicts monthly sentiment trends across the three regions (i.e., DRC, China, and U.S.A. and Canada) over a 12-month period centred around a truckers' strike event in the DRC on the 7th of November 2023. The analysis presents two approaches: conducting the sentiment analysis with and without news article summaries as the context, tracked from six months before to six months after the event. The monthly sentiment was calculated using a one-month rolling window, starting from the event date. For example, +1 months reflects the average sentiment of news published between the 7th of November and the 7th of December 2023, while +2 months corresponds to news published between the 7th of December 2023 and the 7th of January 2024.

Temporal dynamics highlight the evolving nature of sentiment around the event. In the DRC, sentiment scores exhibit an increase approximately one month before the event, with several positive news items like the development of transportation corridors, followed by a substantial and sustained

decline in the post-event period. Meanwhile, sentiment trends in news coverage from China, and the U.S.A. and Canada, remain relatively steady, with only slight negative shifts after the event. The pre-event period showed relatively stable patterns across all regions, while the post-event period revealed increasing divergence, particularly in the DRC's sentiment trajectory compared to other regions. This suggests that the truckers' strike had a lasting negative impact on local sentiment in the DRC while having limited influence on sentiment in other regions.

The results also reveal notable regional variations in sentiment trends. News coverage from the DRC demonstrates considerable volatility, with sentiment scores showing a noticeable decline from positive to negative following the event. China's sentiment trajectory displayed moderate variability within the analysed timeframe, generally maintaining a neutral level, with slightly more stable patterns than the DRC but more fluctuation than the U.S.A. and Canada regions. All these observed patterns are expected, as the truckers' strike occurred in the DRC, directly disrupting local logistics and cobalt shipments. China

was moderately affected as the immediate echelon in the cobalt SC and the primary processor of DRC's minerals. In contrast, as downstream consumers, the U.S.A. and Canada experienced less immediate impact in this case.

The comparison between context-based and context-free sentiment analysis demonstrates the enhanced sensitivity of context-aware approaches in detecting sentiment fluctuations related to the DRC truckers' strikes. The context-based analysis exhibited greater variance in sentiment values across all regions, which is particularly evident in the DRC. This increased sensitivity is most observable during critical periods, such as the sharp sentiment spike one month before the event and the subsequent decline in the DRC. Although both methodologies followed similar directional trends, the context-enhanced approach seemed more effective at detecting subtle shifts in news sentiment, even when aggregating sentiment over a period, which typically smooths out smaller changes. This offers a more detailed understanding of how the strikes' impacts were perceived across different regions.