2026-01-0471 Published 07 Apr 2026

A Large Language Model-Based Database for Analyzing the Battery Critical Minerals Supply Chain

Juntong Zhu, Wei Luo, Xiang Zhang, Zhifeng Yang, and Shiqi(Shawn) Ou South China University of Technology

Xin He Aramco Americas

Citation: Zhu, J., Luo, W., Zhang, X., Yang, Z. et al., "A Large Language Model-Based Database for Analyzing the Battery Critical Minerals Supply Chain," SAE Technical Paper 2026-01-0471, 2026, doi:10.4271/2026-01-0471.

Received: 23 Oct 2025 Revised: 17 Dec 2025 Accepted: 26 Jan 2026

Abstract

Global geopolitical volatility is recognized as a critical threat to the resilience of the electric vehicle battery supply chain. Static, manually updated databases are inadequate for capturing the sector's rapid dynamics, resulting in significant information gaps for strategic planning. To address this, an Artificial Intelligence-driven methodology is proposed for constructing a comprehensive and dynamic database. An automated pipeline was implemented. First, real-time textual data are collected from curated news and industry sources using specialized web crawlers. Then, the unstructured data obtained undergo preprocessing, including deduplication and cleansing, to ensure quality. A core innovation involves the application of Large Language Models (LLMs) for deep semantic parsing and extraction of structured information. These models are utilized to accurately identify key entities—such as corporations, facilities, and production capacities—and to delineate complex multi-tier relationships spanning from raw material extraction to final distribution. The output is a structured database that provides a data-rich representation of the global supply chain. Experimental results demonstrate that the proposed semantic deduplication framework achieves a recall of 86.3% in identifying duplicate content across multilingual texts, significantly outperforming traditional methods. Through this system, over 200,000 news and industry reports have been successfully processed and structured, encompassing more than 5,000 companies worldwide. This approach highlights the transformative potential of LLMs in industrial intelligence, offering a critical tool for enhancing visibility, fostering resilience, and enabling data-driven decision-making for sustainable mobility amid global disruptions.

Introduction

With the rapid expansion of the electric vehicle (EV) industry, its global market share has continued to grow steadily. Taking the Chinese market as an example, the penetration rate of EVs had reached 48% by 2024 [1]. This surge in EV sales consequently led to a substantial increase in the demand for lithium batteries and their raw materials. At the same time, external factors such as international political tensions, geopolitical conflicts, and unforeseen global events have exerted significant influence on the lithium battery materials supply chain [2], resulting in growing systemic complexity. This complexity poses severe challenges for corporate industrial management and the international business performance (IBP) of multinational enterprises, as inadequate management may result in substantial adverse outcomes [3]. More critically, supply chain risks associated with battery materials are prevalent across all stages of production—from mining and smelting to manufacturing [4]—imposing higher demands on firms' supply chain management and risk control capabilities.

In an increasingly turbulent global environment, supply chain resilience has become a critical organizational capability [5]. Traditional research has focused on minimizing vulnerabilities—through redundancy, diversification, or multi-sourcing—or on enhancing agility and contingency planning. To overcome the limitations of these approaches, various artificial intelligence techniques such as machine learning, deep learning, and knowledge graphs have been applied, enabling more comprehensive modeling, prediction, and visualization of supply chain risks [6]. However, these methods often rely heavily on structured internal data and conventional quantitative metrics, struggling to deliver timely and context-aware risk detection when faced with unstructured, open-source information like news articles, policy updates, and market insights.

In recent years, text mining and natural language processing (NLP) techniques have been increasingly explored to extract actionable insights from such unstructured supply chain data [7]. Early studies employed methods like TF-IDF, LDA, and word2vec to identify risk

signals. At the same time, named entity recognition and relation extraction were used to identify key entities such as suppliers and disruptive events. Still, these traditional text mining methods rely heavily on manual feature engineering and struggle to generalize across languages or capture complex semantic relationships. In contrast, LLMs such as GPT, BERT, and LLaMA have demonstrated remarkable capabilities in semantic understanding, natural language generation, and cross-lingual reasoning [8]. These models can perform advanced tasks, including cross-lingual information extraction, event summarization, and knowledge graph construction, offering flexibility and scalability beyond traditional NLP [9]. Despite their adoption in domains like customer service and demand forecasting, the application of LLMs in supply chain intelligence and risk analysis remains largely unexplored. Few studies have systematically integrated LLMs to extract actionable knowledge from unstructured, open-source supply chain data, presenting both a challenge and an opportunity.

The advances in NLP and LLMs now make it feasible to build a dynamic and reliable supply chain intelligence database from open-source online news [10]. However, realizing this potential for the battery materials industry presents several critical challenges: first, the multi-source, multi-language, and multi-format nature of news leads to fragmented and inconsistent data, complicating unified information extraction; second, the high redundancy of news content due to widespread reposting and cross-platform citation introduces significant noise; and third, the inconsistent credibility of sources risks incorporating biased or false information, which could mislead decision-making. To overcome these challenges, this study develops a comprehensive and reliable pipeline for extracting structured supply chain information. The main contributions of this work are as follows:

1. **An end-to-end intelligent framework for supply chain news analysis.** Focusing on the battery materials domain, this study proposes an innovative global supply chain news processing framework that integrates data acquisition, semantic modeling, deduplication, and credibility assessment. This framework is designed to support multiple data sources and efficiently handle heterogeneous supply chain-related information, directly addressing the challenge of data fragmentation and inconsistency.
2. **Construction of an open-source news database.** A large-scale bilingual (Chinese-English) open-source news database has been developed, containing over 100,000 records. Beyond battery material supply chains, the database encompasses policy developments, market trends, and industrial dynamics. This structured repository mitigates content redundancy by integrating deduplication and provides comprehensive support for supply chain management and research.
3. **System validation and experimental evaluation.** Extensive experiments were conducted to validate the performance of key system modules, including information extraction accuracy, redundancy elimination, and credibility scoring. Results demonstrate that the optimized system achieves improved precision, deduplication effectiveness, and reliability assessment. This ensures the quality of the extracted intelligence and indicates strong potential for real-world applications in the battery materials industry.

The remainder of this paper is organized as follows. Section 2 presents the methodological framework in detail. Section 3 reports experimental evaluations of system modules. Section 4 discusses the advantages, limitations, and potential applications of the proposed system. Finally, Section 5 concludes the paper and outlines directions for future work.

Methodology

The proposed system is an end-to-end intelligent framework for supply chain information extraction and risk monitoring, leveraging open-source news data. Figure 1 presents the system's flowchart. The system consists of six core modules: Crawler, Preprocessing, Vectorization, Similarity Filtering, Credibility Scoring, and Storage/Query. Each module is functionally independent yet seamlessly integrated through standardized data interfaces to ensure efficiency and scalability.

Data Acquisition and Cleaning

In this study, the data acquisition module serves as the foundational component for continuously collecting open-source supply chain news related to battery materials from multiple sources. To ensure data diversity, the selected news sources cover different languages, websites, and platforms, including but not limited to mainstream media outlets, specialized battery industry news portals, and official industry announcements. To ensure timely information and better support enterprise-level supply chain management and decision-making, the module is designed with a dynamic incremental acquisition mechanism. It operates in an automated manner, continuously monitoring relevant platforms for newly released supply chain news, retrieving qualified articles, and transferring them to the data preprocessing module via a standardized interface. Figure 2 presents the detailed process of data processing.

The data preprocessing module primarily performs two key functions: format transformation and data cleaning. Given the high heterogeneity of news sources and formats, it is essential to first convert various raw data formats into a unified structure suitable for downstream processing. During the data cleaning phase, a series of automated filtering and refinement procedures is applied to remove low-quality entries. The system first identifies the main body of each article and eliminates

FIGURE 1 System Flowchart

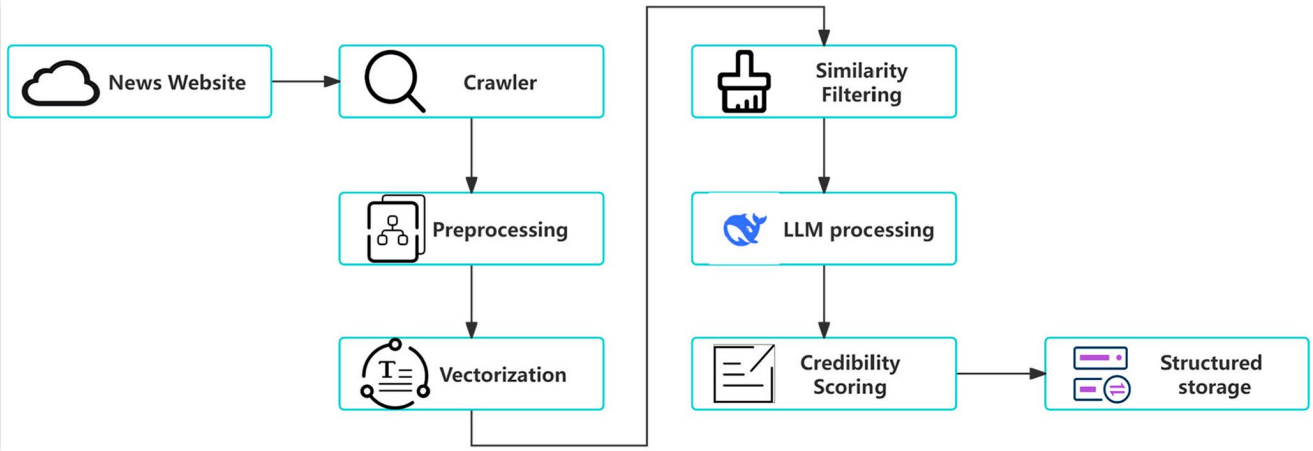
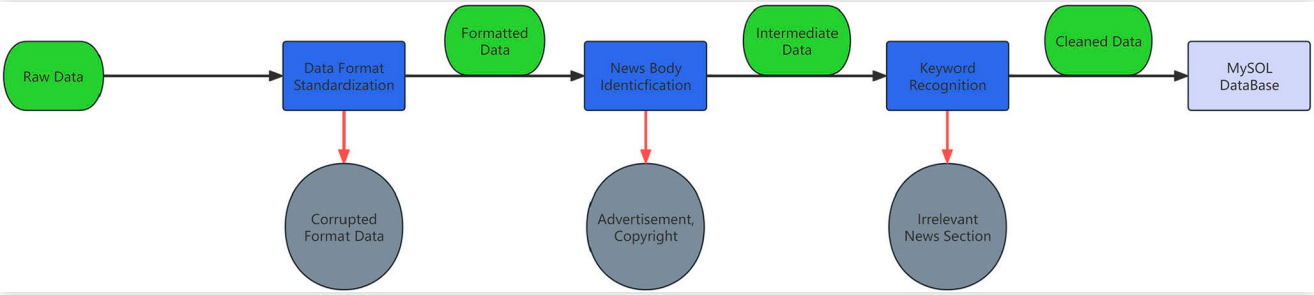


FIGURE 2 Data Processing Flowchart



irrelevant elements such as advertisements and embedded watermarks. Subsequently, a keyword-based filtering mechanism is implemented to exclude content that is clearly unrelated to the battery materials supply chain, ensuring the overall reliability and domain relevance of the dataset.

Finally, all processed news data are stored in a structured database, forming a solid foundation for subsequent semantic modeling, deduplication, and credibility assessment. Through this automated data collection and preprocessing workflow, the proposed system ensures that high-quality, relevant, and timely news data are efficiently acquired and made readily available for downstream analytical modules, thereby enhancing the system's overall reliability and responsiveness.

Data Preprocessing and Semantic Modeling

To further enhance the quality of the preprocessed dataset before semantic deduplication, the system incorporates a blacklist-based filtering mechanism. During preprocessing, instances of irrelevant or low-quality content—such as advertisements, watermarks, and malformed text—are automatically detected and added to a dynamic blocklist. Once identified, any subsequent occurrences of these elements are filtered out in real time, preventing them from entering the downstream processing pipeline. This approach not only maintains the integrity and reliability but also mitigates noise that could adversely affect semantic embedding and deduplication performance.

Building upon the data preprocessing stage, this study develops a domain-specific multi-level keyword system tailored for the battery materials supply chain. The purpose of this system is to accurately identify and extract analytically valuable information from vast amounts of unstructured news data. The keyword system adopts a three-tier hierarchical architecture, designed to comprehensively capture the full spectrum of terminology used across the supply chain.

At the foundational terminology level, the system encompasses essential vocabulary spanning all stages of the battery materials industry chain. This includes upstream resources such as key minerals (e.g., lithium, cobalt, nickel); midstream materials such as cathode materials (e.g., Nickel Cobalt Manganese battery materials, Lithium Iron Phosphate battery materials), anode materials (e.g., synthetic graphite, silicon-carbon composites),

electrolytes, and separators; and downstream products such as power batteries, energy storage systems, and EVs.

The enterprise terminology level focuses on the inclusion of major global companies across different segments of the industry, including mining enterprises, materials manufacturers, battery producers, and vehicle assemblers.

The supply chain terminology level systematically integrates key operational indicators and business metrics throughout the entire supply chain process. These include capacity, output, pricing, inventory, procurement, delivery, and capacity utilization, among others.

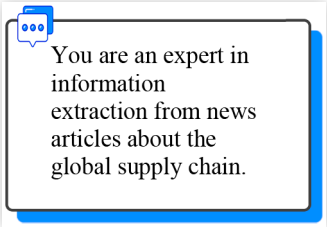
Based on this comprehensive keyword system, the study further establishes a multi-dimensional content value evaluation model. This model assesses each news article through an integrated analysis of multiple dimensions—such as the coverage of industry-chain keywords, the density of domain-specific terminology, and the completeness of supply chain indicators. By assigning a precise value score to each piece of text, the system ensures that only high-quality news with substantive industrial information advances to subsequent analytical modules. This mechanism guarantees data quality at the source, thereby significantly enhancing the system's accuracy in analyzing the battery materials supply chain and its capability to support informed decision-making.

Prompt Design

In the context of LLMs, prompt design plays a central role in guiding the model to perform downstream tasks efficiently. Unlike traditional supervised learning systems, which require paired input-output data for training, LLMs leverage their pre-trained knowledge to model the probability of text sequences directly. By carefully crafting a prompt, a downstream task can be reformulated as a cloze-style or text-generation problem, which aligns with the objective the model was originally trained to solve on large-scale corpora. A well-designed prompt thus acts as an instructive context, effectively bridging the gap between the model's pre-training and the desired task, allowing the LLM to produce accurate, context-aware outputs without extensive task-specific parameter updates [11]. This capability is particularly valuable in domains like supply chain intelligence, where labeled datasets are scarce and task-specific annotations are expensive to obtain [12].

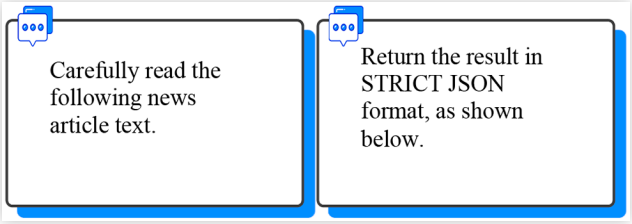
1. System Prompt The LLM is configured to identify the role of each company within the supply chain, distinguishing between production and transactional activities, and to enforce a standardized output format for consistency [13]. Based on the assigned role, the system populates relevant domain-specific information while ensuring that all outputs are presented in English. For missing or incomplete data, specialized handling is applied: entries lacking critical information are considered invalid and removed from the dataset. This approach guarantees that only high-quality, structured, and meaningful supply chain information is retained, providing a reliable foundation for subsequent analysis and decision-making.

FIGURE 3 System Prompt Example



2. User Prompt The user prompt defines the LLM's professional role as an expert in extracting supply chain information from news sources. It specifies the task of accurately identifying and extracting relevant relationships from the text [14]. All outputs must adhere to a predefined structured format, distinguishing between production/processing and transactional relationships, and populating role-specific information accordingly. Standardized conventions, such as English-language output and consistent material naming, are enforced, and entries lacking valid information are omitted. To enhance performance, the prompt includes illustrative examples from multilingual sources, enabling the model to learn both the expected extraction patterns and formatting rules, thereby improving its ability to generate consistent, high-quality structured data [15].

FIGURE 4 User Prompt Example

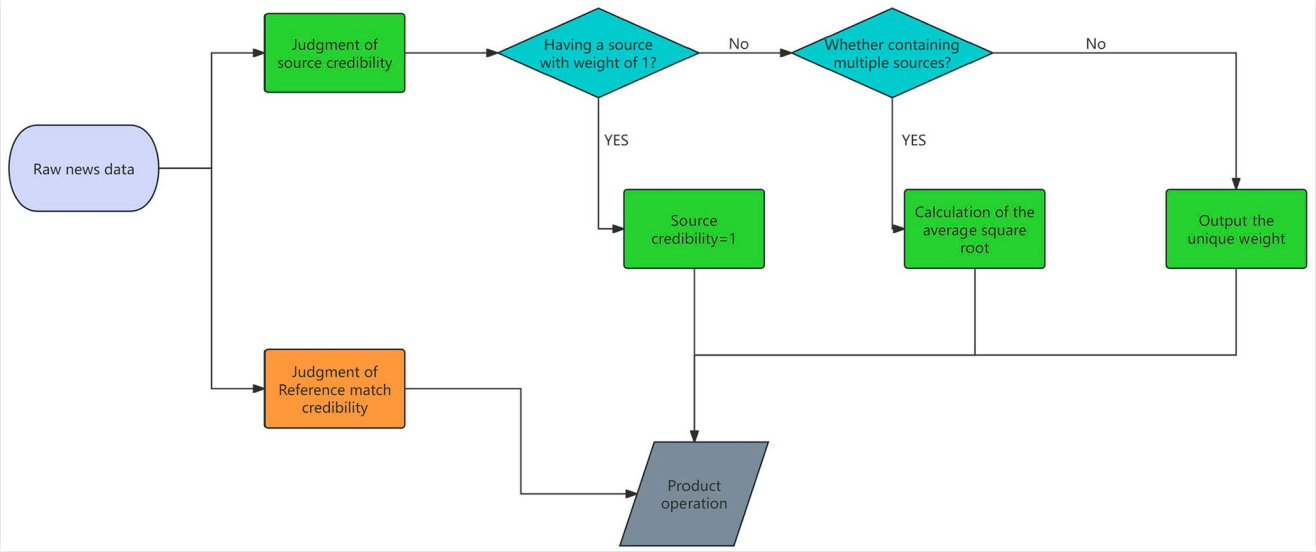


Credibility Assessment

Given the proliferation of false and misleading news online, evaluating the credibility of extracted information is as crucial as the extraction itself. Erroneous or unverified data can lead to significant analytical and operational losses [16]. Traditional manual fact-checking is insufficient for large-scale news processing; therefore, this study integrates an automated, interpretable, rule-based credibility assessment into the extraction workflow, assigning a quantitative credibility score to each supply chain relationship. The details of the credibility assessment process are shown in Figure 5.

The evaluation is operationalized along two complementary dimensions: Source Credibility and Content Credibility. Source Credibility reflects the authority and historical reliability of the news outlet, with official releases or enterprise reports rated highest, unsourced information given a neutral baseline, and unverified industry rumors rated lower.

FIGURE 5 Credibility Assessment Flowchart



Content Credibility assesses the degree to which extracted information is explicit, factual, and clearly integrated within the source. High scores are assigned to concrete, verifiable statements, while ongoing projects, plans, indirect reports, and forward-looking technological statements receive progressively lower scores reflecting increasing uncertainty. A default neutral score is applied when explicit linguistic or contextual cues are absent.

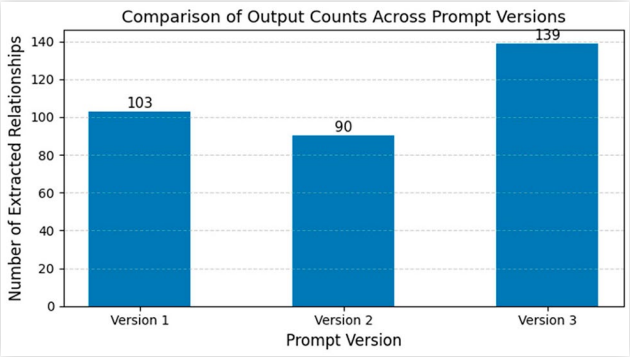
This multidimensional scoring framework ensures that the system prioritizes reliable and verifiable information, providing a robust foundation for downstream supply chain analysis and decision-making.

Result

Prompt Engineering Experiment

To investigate the impact of prompt design on information extraction performance, this study conducted a comparative experiment across three progressively refined prompt versions. Figure 6 shows the results of different prompts.

FIGURE 6 Comparison of Prompt



Version 1 provided only the task description and basic extraction rules, instructing the LLM to identify and structure supply chain relationships from news articles.

Version 2 extended this by introducing an explicit constraint — FORMAT TO FOLLOW — which defined the expected JSON schema to ensure standardized outputs.

Version 3 further incorporated a concrete example of a news article and its corresponding extraction result, enabling the model to align its reasoning with a clear demonstration of the desired behavior.

All three versions were evaluated on the same set of 100 English news articles. The results demonstrate a consistent pattern: Version 1 produced 103 extracted relationships, Version 2 generated 90, and Version 3 yielded 139. The differences reflect distinct trade-offs between output completeness and structural compliance. Version 1 tended to create an excess of marginally relevant outputs, while Version 2's stricter formatting constraint improved consistency but occasionally suppressed borderline valid cases. Version 3, benefiting from in-context learning, achieved both higher coverage and improved structural accuracy by grounding the model's generation in a concrete, task-specific example.

These findings highlight that effective prompt design—particularly the inclusion of explicit format guidance and illustrative examples—plays a crucial role in enhancing LLM performance for domain-specific information extraction tasks. This underscores the necessity of iterative prompt optimization as an integral part of system development.

Semantic-Based News Deduplication Method

Building upon the precise identification of high-value content, this study introduces an innovative deep learning–based semantic deduplication framework

specifically designed to address the challenges of multilanguage and multi-expression redundancy in battery materials supply chain news. Given the characteristics of such texts—frequent code-switching between languages, dense technical terminology, and diverse semantic representations—an extensive study was conducted. The paraphrase-multilingual-MiniLM-L12-v2 model within the Sentence-BERT framework was selected as the core semantic understanding engine.

This model demonstrates exceptional multilingual comprehension by accurately interpreting mixed Chinese–English professional texts. Its pretraining on large-scale corpora enables a deep contextual understanding of semantics relevant to the battery materials domain. In practical application, the system first performs fine-grained preprocessing on news articles that have passed the content value assessment stage. This includes sentence segmentation and semantic unit division, ensuring precise granularity for embedding generation. The processed text is then fed into the model to generate high-dimensional sentence embeddings.

By computing the cosine similarity among these embeddings, the system constructs a semantic correlation matrix that captures deep-level meaning relationships between news texts. This enables the detection of semantically redundant information—articles that differ in surface phrasing yet convey essentially identical content.

To empirically evaluate the model's performance, experiments were conducted on 500 pairs of news samples, among which 73 pairs contained actual duplicate content. For comparative analysis, two baseline methods were implemented: the TF-IDF frequency-based approach and the simHash locality-sensitive hashing method. The results, visualized through confusion matrices in Figures 7–9, demonstrate that the proposed semantic deduplication framework achieves superior accuracy and robustness in identifying content-level redundancy across multilingual and domain-specific contexts.

FIGURE 7 TF-IDF Confusion Matrix

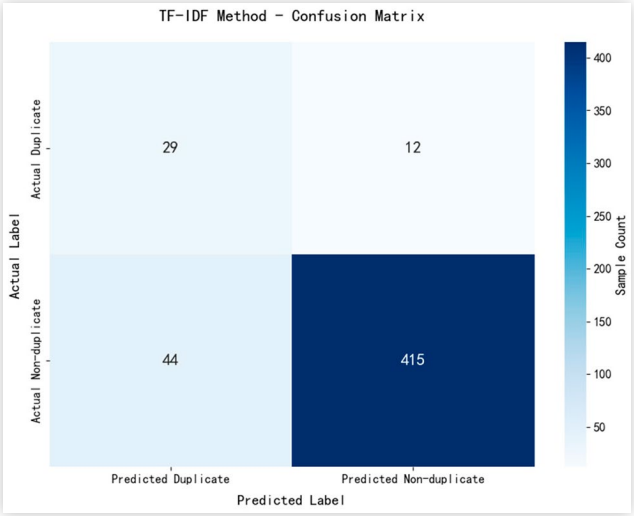


FIGURE 8 SimHash Confusion Matrix

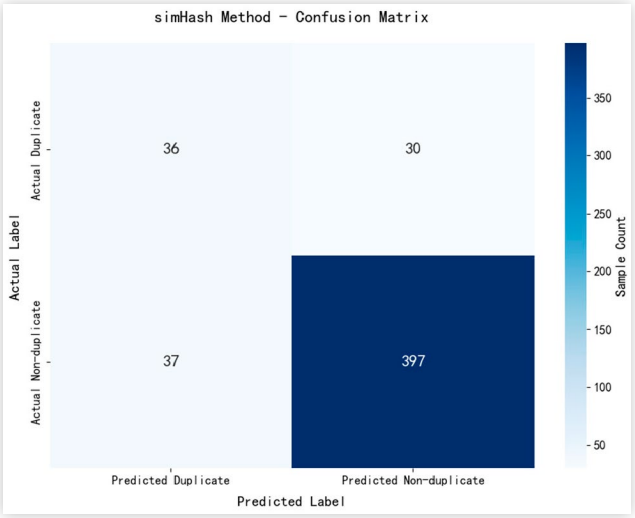
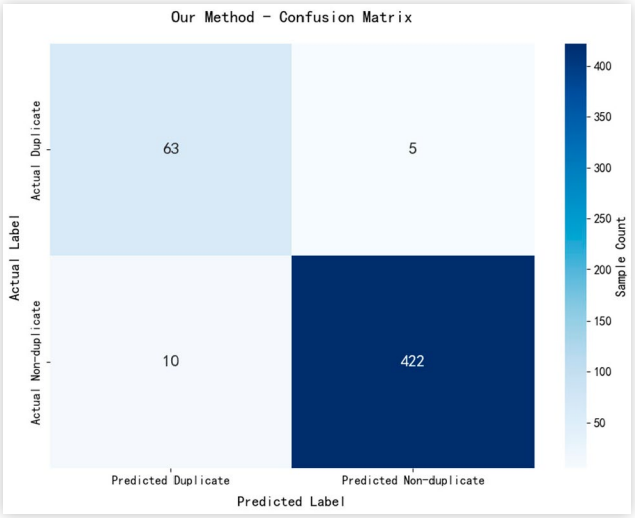


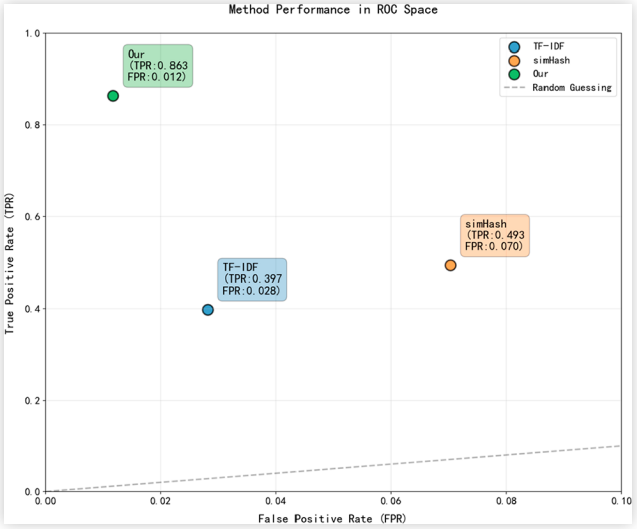
FIGURE 9 Our Method



In terms of performance, the proposed method demonstrates significant advantages over traditional approaches. As illustrated in Figure 10, this semantic deduplication framework consistently outperforms both the TF-IDF and simHash methods across all key evaluation metrics. In particular, the recall rate of this model reaches 86.3%, far exceeding that of TF-IDF (39.7%) and simHash (49.3%). This clearly indicates that the semantic understanding–based deduplication approach can more comprehensively capture duplicated content, effectively reducing the omission of valuable information.

Furthermore, this method achieves an excellent precision rate of 92.6%, substantially higher than TF-IDF (70.7%) and simHash (54.5%). This result highlights the model's strong capability to accurately identify redundant information while minimizing false positives. Notably, the false positive rate of the approach is only 1.1%, significantly lower than that of TF-IDF (2.8%) and simHash (7.0%). This low error rate is particularly crucial in real-world

FIGURE 10 ROC Spatial Performance map



applications, as it ensures that critical information is preserved without being mistakenly filtered out.

From an overall perspective, the method achieves an F1 score of 0.894, reflecting an excellent balance between precision and recall. This superior performance can be attributed to the semantic representation power of the Sentence-BERT model, which captures the deep contextual meaning of text rather than relying solely on surface-level lexical similarity. Specifically, while TF-IDF can detect basic lexical overlaps, it fails to recognize semantic equivalence across different wordings, leading to a high rate of missed detections. Similarly, although simHash provides computational efficiency, its limited sensitivity to nuanced semantic differences results in frequent false matches.

In contrast, this model-generated semantic embeddings effectively encode the intrinsic meaning of supply chain–related texts. When combined with a domain-optimized similarity threshold, this enables the system to achieve the optimal trade-off between accuracy and coverage. Consequently, this deep semantic deduplication method is particularly well-suited for handling multilingual, multi-source, and variably phrased professional news in the battery materials supply chain domain, providing a far more reliable data foundation for subsequent analytics and decision-support processes.

Implementation Details and Resource Requirements

To assess the practical feasibility and scalability of the proposed system for industrial deployment, a quantitative analysis of its computational efficiency and associated costs was conducted. Understanding the processing time and financial expenditure per article is crucial for evaluating whether the system can sustainably handle the real-time, large-volume data streams typical of global supply chain monitoring.

A controlled experiment was performed using a representative sample of 1,000 English and 1,000 Chinese news articles. The processing pipeline (including crawling, preprocessing, and deduplication) is executed on a local server. The core LLM-based information extraction stage utilizes the DeepSeek-V3.2-chat API for structured data parsing. The detailed performance and cost results are summarized in Table 1.

The results highlight two key points. First, processing Chinese news requires approximately 2.3 times more time and 1.96 times higher cost than English news for the sample, primarily due to the greater average text length and the tokenization mechanism of the LLM. Second, the monetary cost remains relatively low at scale; processing 200,000 articles (as reported in this study) is estimated to incur inference costs between CNY 552 and CNY 1,080, demonstrating the cost-effectiveness of the pipeline. The system architecture supports batch processing and parallel API calls, allowing for linear scaling with additional computational resources, which is a critical feature for enterprise-level applications.

Discussion

The proposed system demonstrates several notable advantages. First, it provides an end-to-end solution for extracting, structuring, and assessing supply chain information from heterogeneous news sources, integrating data acquisition, preprocessing, semantic deduplication, and credibility evaluation within a unified framework. Second, the system demonstrates cross-linguistic capability, effectively processing both Chinese and English news articles, thereby broadening its applicability across global supply chains. Third, the architecture is designed for large-scale scalability, enabling the automated handling of extensive news volumes while maintaining data quality and analytical consistency. These characteristics collectively support more efficient supply chain monitoring, trend analysis, and risk assessment, offering enterprises a practical tool for informed decision-making.

Despite these strengths, the system has several limitations. Currently, the implementation primarily focuses on Chinese and English sources, leaving the performance on additional languages untested. Furthermore, while the credibility scoring mechanism provides a quantitative

TABLE 1 Computational Performance and Cost Analysis for News Processing

Metric	Average Length	Avg. Processing Time per Article	Total Token Consumption	Total Estimated Cost (API Call)
English News (n=1,000)	308.07 words	7.67 s	5,882,185	CNY 2.76
Chinese News (n=1,000)	1,007.79 characters	17.79 s	8,000,513	CNY 5.40

measure of news reliability, its effectiveness requires further empirical validation, especially in rapidly evolving or less standardized media environments. To address the need for empirical validation of the credibility module, a comprehensive evaluation is planned for future work, which will involve manual auditing and cross-validation using large language models to calculate accuracy and misjudgment rates. This will provide a quantitative assessment of the credibility module's performance. In addition, the system relies on textual news data; integrating other unstructured or semi-structured sources, such as social media or regulatory filings, may pose additional challenges. These limitations highlight opportunities for refinement and future enhancement.

The system's potential applications are extensive. It can support supply chain intelligence, including identification of production and transactional relationships, monitoring of material flows, and early detection of supply disruptions. Moreover, by aggregating structured information over time, it can facilitate trend analysis and risk forecasting, enabling enterprises to anticipate market fluctuations and strategically adjust sourcing or production plans.

Conclusion and Future Work

This study presents a cross-lingual, scalable news processing system for the battery materials supply chain, integrating automated information extraction, semantic deduplication, and credibility evaluation. Experimental results demonstrate that the proposed system effectively extracts relevant relationships, eliminates redundant content, and assigns meaningful reliability scores. Specifically, the system implements an end-to-end framework capable of automatically collecting news articles from multiple web sources and transforming them into structured data via a large language model. To date, this framework has processed over 200,000 original news entries, thereby establishing a robust foundation for data-driven supply chain analysis.

Future work aims to enhance the system's analytical capabilities. First, integrating a knowledge graph will enable the construction of a comprehensive network of supply chain relationships, supporting more sophisticated queries and visualizations. Second, incorporating causal inference and predictive modeling will allow the system to anticipate potential supply chain risks and forecast market trends, moving beyond descriptive analytics toward proactive decision support. Finally, expanding the system to accommodate additional languages and heterogeneous data sources will further strengthen its global applicability, ensuring that enterprises can rely on it for comprehensive, timely, and actionable supply chain intelligence.

References

1. International Energy Agency, "*Global EV Outlook 2025*," 2025.
2. Bednarski, L., Roscoe, S., Blome, C., and Schleper, M.C., "Geopolitical Disruptions in Global Supply Chains: A State-of-the-Art Literature Review," *Production Planning & Control* (2023): 1-27, doi:10.1080/09537287.2023.2286283.
3. Sharma, A., Kumar, V., Borah, S.B., and Adhikary, A., "Complexity in a Multinational Enterprise's Global Supply Chain and Its International Business Performance: A Bane or a Boon?" *Journal of International Business Studies* 53, no. 5 (2022): 850-878, doi:10.1057/s41267-021-00497-0.
4. Sun, X., Hao, H., Hartmann, P., Liu, Z. et al., "Supply Risks of Lithium-Ion Battery Materials: An Entire Supply Chain Estimation," *Materials Today Energy* 14 (2019): 100347, doi:10.1016/j.mtener.2019.100347.
5. Gupta, S., Modgil, S., Meissonier, R., and Dwivedi, Y.K., "Artificial Intelligence and Information System Resilience to Cope with Supply Chain Disruption," *IEEE Transactions on Engineering Management* 71 (2024): 10496-10506, doi:10.1109/TEM.2021.3116770.
6. Pettit, T.J., Fiksel, J., and Croxton, K.L., "Ensuring Supply Chain Resilience: Development of a Conceptual Framework," *Journal of Business Logistics* 31, no. 1 (2010): 1-21.
7. Gupta, V. and Lehal, G.S., "A Survey of Text Mining Techniques and Applications," *Journal of Emerging Technologies in Web Intelligence* 1, no. 1 (2009): 60-76.
8. Yao, Y. et al., "A Survey on Large Language Model (LLM) Security and Privacy: The Good, the Bad, and the Ugly," *High-Confidence Computing* 4, no. 2 (2024): 100211.
9. Ge, Y. et al., "Openagi: When LLM Meets Domain Experts," *Advances in Neural Information Processing Systems* 36 (2023): 5539-5568.
10. Bi, X. et al., "Deepseek LLM: Scaling Open-Source Language Models with Longtermism," *arXiv preprint* arXiv:2401.02954, 2024.
11. Shin, T. et al., "AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts," *arXiv preprint* arXiv:2010.15980, 2020.
12. Love, M., "Enhancing Supply Chain Through Prompt-Language Model," unpublished manuscript, ResearchGate, 2025.
13. Gu, J. et al., "A Survey on LLM-as-a-Judge," *arXiv preprint* arXiv:2411.15594, 2024.

14. Zheng, L. et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," *Advances in Neural Information Processing Systems* 36 (2023): 46595-46623.
15. Wu, Q. et al., "Autogen: Enabling Next-Gen LLM Applications via Multi-Agent Conversations," *First Conference on Language Modeling* (2024).
16. Liu, J., Zhang, L., Munir, S., Gu, Y. et al., "VeriFact: Enhancing Long-Form Factuality Evaluation with Refined Fact Extraction and Reference Facts," *arXiv preprint* arXiv:2505.09701, 2025.

Contact Information

Shiqi(Shawn) Ou
sou@scut.edu.cn; oushiqi@pazhoulab.cn
Phone number: +86-020-81181684

Mailing address:
1. South China University of Technology, School of Future Technology, 777 Xingye Ave East, Panyu District, Guangzhou, Guangdong, 511442, China
2. Guangdong Artificial Intelligence and Digital Economy Laboratory (Guangzhou), 70 Yuean Road, Haizhou District, Guangzhou, Guangdong 510335, China.

Acknowledgments

This work is funded by the Development Agreement between Saudi Aramco Technologies Company and South China University of Technology (SCUT) (SATC 2025-024) and used the facilities of the Pazhou Laboratory. All opinions expressed in this paper are the authors' and do not necessarily reflect the policies and views of SCUT, Pazhou Laboratory, or Saudi Aramco.

© 2026 SAE International. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, or used for text and data mining, AI training, or similar technologies, without the prior written permission of SAE.

Positions and opinions advanced in this work are those of the author(s) and not necessarily those of SAE International. Responsibility for the content of the work lies solely with the author(s).

ISSN 0148-7191