Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16477
e-ISSN : 2541-2019
p-ISSN : 2541-044X

SC-Literature Intelligence: A Retrieval-Augmented Generation Framework for Multi-Category AI Literature Synthesis in Supply Chain

Setio Basuki[1)*], Amelia Khoidir[2)], Ilham Perdana[3)], Muhammad Daffa Nugraha[4)], Masatoshi Tsuchiya[5)]

[1,3,4)]Department of Informatics, Universitas Muhammadiyah Malang, Indonesia
[2)]Industrial Engineering Department, Universitas Muhammadiyah Malang, Indonesia
[5)]Department of Computer Science and Engineering, Toyohashi University of Technology, Toyohashi City, Japan
[1)]setio_basuki@umm.ac.id, [2)]ameliakhoidir@umm.ac.id, [3)]ilhamperdana@umm.ac.id,
[4)]nugrahadaffa568@webmail.umm.ac.id , [5)]tsuchiya@imc.tut.ac.jp

Submitted : June 26, 2026 | **Accepted** : July 4, 2026 | **Published** : July 6, 2026

Abstract: This paper develops SC-Literature Intelligence, a retrieval-augmented generation (RAG) framework for research synthesis of scientific literature on artificial intelligence (AI) in the supply chain domain. The study addresses the fragmentation of scientific findings, which makes cross-document understanding difficult by supporting four categories of literature-analysis queries: trend analysis, gap detection, comparative synthesis, and evidence-based question answering (QA). The primary novelty lies in introducing a category-aware research synthesis framework capable of evaluating RAG performance across multiple literature-analysis tasks rather than conventional question answering. The framework is built from Scopus-indexed abstracts through pre-processing, chunk-based embedding using BGE-M3 and LaBSE, vector storage, semantic retrieval, and prompt-guided generation evaluated using the RAGAS framework across 640 experimental runs. The results show that BGE-M3 consistently outperforms LaBSE on all RAGAS indicators with the best configuration (chunk size 64, Top-K 5) achieving scores between 0.722 and 0.856 across faithfulness, answer relevancy, context precision, and context recall. Gap detection emerges as the best-supported query category, whereas comparative synthesis remains the most challenging. Failure analysis further reveals that retrieval-stage issues dominate over generation-stage issues, identifying embedding quality as the primary bottleneck. These findings demonstrate that category-aware RAG-based synthesis can support structured, evidence-grounded literature analysis in the supply chain AI domain.

Keywords: retrieval-augmented generation, large language model, research synthesis, supply chain, embedding model, literature analysis.

INTRODUCTION

Research in the supply chain field is experiencing rapid development along with the increasing adoption of artificial intelligence (AI) to support complex decision making (Daios, Kladovasilakis, Kelemis, & Kostavelis, 2025). AI has been used in various supply chain contexts, from demand forecasting (Douaioui, Oucheikh, Benmoussa, & Mabrouki, 2024), risk management (Shamsuddoha, Khan, Chowdhury, & Nasir, 2025), disruption mitigation (Riad, Naimi, & Okar, 2024), to supply chain resilience (Dindarian, 2026). Report in 2026 (Storan Dylanand Chiarelli, 2026), there are 5.7 million articles, reviews, and conference papers in 2024. This explosion in the number of publications is also reflected in the Compound Annual Growth Rate (CAGR) from 2014 to 2024 which is 4%. This situation makes it difficult for researchers to gain a comprehensive understanding of research trends (Abogunrin et al., 2025), dominant methodological approaches, and the real contribution of AI to the supply chain. The process of literature exploration becomes complex because

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

scientific information is spread in fragmented and huge numbers of papers. This opens opportunities for technology, especially AI to contribute to simplifying the research synthesis process.

Although research synthesis tools such as Scopus, Web of Science, Semantic Scholar, PubMed, and Scite.ai are useful for literature search, they have clear limitations. They return lists of relevant documents but cannot perform cross-document reasoning (Bolaños, Salatino, Osborne, & Motta, 2024). In addition, topic modeling-based document analysis methods such as latent dirichlet allocation (LDA) (Galán-Mena et al., 2026) and Bidirectional Encoder Representations from Transformers (BERTopic) (George & Sumathy, 2023) produce static representations in the form of keyword lists that are difficult to interpret (Hoyle et al., 2021). The manual synthesis process using results obtained from literature search technology tends to be non-scalable, time-consuming, and potentially biased (Elliott & Turner, 2022). Some of these weaknesses are caused by variations in the background, experience and references of researchers (Wagner, Lukyanenko, & Paré, 2022). Taken together, the documented inability of retrieval platforms to reason across documents (Bolaños et al., 2024), the interpretability limits of static topic representations (Hoyle et al., 2021), and the scalability and bias problems of manual synthesis (Elliott & Turner, 2022; Wagner et al., 2022) indicate that current technological approaches do not yet support systematic and automated research synthesis.

The advancement of large language models (LLMs) brings opportunities to automate research synthesis (Bolaños et al., 2024). LLMs exhibit strong contextual modeling capabilities by enabling coherent text generation, structured summarization, and supporting higher-level analytical tasks (Sajjadi Mohammadabadi et al., 2025). Despite these advantages, using LLMs without authoritative external references can pose a risk of hallucinations (L. Huang et al., 2025). Hallucination is a phenomenon where LLMs provide answers that linguistically seem convincing and coherent, but are factually inaccurate (Ji et al., 2023). This is a serious issue in the context of scientific research, where reliability and traceability are crucial. To address this, LLMs need to be equipped with a mechanism to support references to factual and authoritative data using Retrieval-Augmented Generation (RAG) (Lewis et al., 2020). In this context, RAG can also act as a cognitive aid that supports structured reasoning and insight generation (Fan et al., 2024).

Based on these gaps, this study is guided by three research questions. RQ1: to what extent do the choice of embedding model and the retrieval configuration (chunk size and Top-K) affect the quality of RAG-based research synthesis as measured by the RAGAS metrics? RQ2: how does synthesis quality vary across the four categories of literature-analysis queries, namely trend analysis, gap detection, comparative synthesis, and evidence-based QA? RQ3: at which stage (retrieval or generation) do failures predominantly occur and what are their dominant types?

To address the gaps identified above, this paper develops SC-Literature Intelligence, a RAG system based on LLMs for research synthesis of AI applications in the supply chain domain. It supports four question categories, trend analysis, gap detection, comparative synthesis, and evidence-based QA. The system runs through five stages, acquiring paper abstracts from Scopus, data pre-processing (cleaning, tokenizing, chunking), building an embedding model and vector database, retrieving Top-K chunks by semantic similarity to augment the LLM, and evaluating with RAGAS. The main contributions of this paper are: (i) a category-aware RAG framework for structured research synthesis in the supply chain AI domain; (ii) a systematic evaluation design that isolates the effects of embedding model, chunk size, and Top-K across the four query categories, answering RQ1 and RQ2; and (iii) a two-stage failure taxonomy that separates retrieval-stage from generation-stage errors and identifies their dominant types, answering RQ3. The quantitative evidence supporting these contributions is presented in the Result and Discussion sections.

LITERATURE REVIEW

RAG-based research has been implemented in various domains, such as healthcare, scientific discovery, and even supply chain. In the healthcare domain, RAG systems emphasize accuracy, reliability, and source traceability. Specifically, research can be grouped into several main focuses: literature review and framework analysis (Amugongo, Mascheroni, Brooks, Doering, & Seidel, 2025), clinical question answering and decision support (Shi et al., n.d.; Zhang et al., 2025; Zhang Shiranand Phan, 2026), electronic health records (EHR)-based summarization and information extraction (Saba, Wendelken, & Shanahan, n.d.), and retrieval optimization methods consist of chunking techniques and domain-specific retrieval enhancements (Hou Yuand Bishop, 2025). Furthermore, in the scientific domain, research focuses on developing innovations from a methodological perspective. Some sub-focuses in this domain include building question answering and research assistant systems (Besrour, He, Schreieder, & Färber, 2025; Hu et al., 2025; Oliva, 2024), improving retrieval mechanisms based on hierarchical retrieval and structure-aware retrieval (Che, Mao, Lan, & Huang, 2024; Gokdemir et al., 2025; Lee, Sohn, Lee, &

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16477
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Kim, 2025; Xu et al., 2025) and literature synthesis and knowledge discovery (Aytar, Kilic, & Kaya, 2024; Godinez, 2025; Krotkov et al., 2025).

In contrast to these two domains, the application of RAG in the supply chain domain is still partial and application-oriented. Most research focuses on decision support and optimization tasks (Wasserkrug, Boussioux and Sun, 2024; Zhu and Vuppalapati, 2024; Li et al., 2025; Simchi-Levi et al., 2025; Zhu Beileiand Vuppalapati, 2025) supply chain mapping and supplier discovery (Jackson, Saénz, Ivanov, & Ma, 2026; Y. Li, Ko, & Ameri, 2025), operational use cases such as warehouse automation and real-time risk detection(Ponnock et al., 2025), and a small number of researches focus on the use of LLMs in supply chain at a conceptual or educational level (Ehrenthal Joachim C. F.and Gachnang, 2024; Huang et al., 2025a).

Table 1 contrasts representative prior works by domain, their reported limitations, and the research gaps they leave open. The comparison shows that earlier systems fall short of research synthesis for two recurring reasons (1) their retrieval and evaluation designs are bound to a single task and (2) their answer quality is never examined across different types of literature-analysis queries, so failures specific to trend, gap, or comparative reasoning remain invisible.

Table 1. Critical Review of Representative RAG Studies and the Remaining Research Gaps.

Study	Domain	Reported Limitation	Research Gap Left Open
Amugongo et al. (2025)	Healthcare	RAG for literature review and framework analysis oriented to clinical accuracy and traceability; single-task design	No structured synthesis across multiple categories of literature-analysis queries
Shi et al. (n.d.); Zhang et al. (2025)	Healthcare	Clinical QA and decision support tuned to factual lookup on curated sources	Evaluation restricted to conventional QA; trend, gap, and comparative analysis untested
Besrour et al. (2025); Hu et al. (2025); Oliva (2024)	Scientific discovery	Research-assistant and QA systems that answer isolated questions	No cross-document, category-aware research synthesis
Che et al. (2024); Lee et al. (2025); Xu et al. (2025)	Scientific discovery	Retrieval-mechanism improvements (hierarchical, structure-aware) benchmarked on generic QA	Answer quality not analyzed across literature-analysis task types
Wasserkrug et al. (2024); Li et al. (2025); Zhu & Vuppalapati (2025)	Supply chain	Decision-support and optimization applications; task-specific, operational focus	No research-synthesis or literature-intelligence objective
Jackson et al. (2026); Y. Li, Ko, & Ameri (2025); Ehrenthal & Gachnang (2024)	Supply chain	Supplier discovery and mapping, or conceptual and educational treatments of LLMs	No system targeting cross-document research intelligence at the synthesis level

Based on the review results, three gaps are identified: (i) existing RAG systems are not designed for structured multi-category research synthesis encompassing trend analysis, gap detection, comparative synthesis, and evidence-based QA nor do they provide systematic analysis of answer quality across these categories; (ii) RAG research is disproportionately concentrated in healthcare and scientific domains, leaving the supply chain domain underserved despite its growing volume of AI literature; (iii) RAG applications in the supply chain domain remain predominantly task-specific and operational, with no system targeting cross-document research intelligence at the synthesis level.

METHOD

The RAG system, based on LLMs for research synthesis, focused on the application of AI in the supply chain domain, is realized through several stages: data acquisition, data pre-processing, building vector embedding, retrieval process, prompt engineering, and evaluation. The system development stages are presented in Figure 1.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16477
e-ISSN : 2541-2019
p-ISSN : 2541-044X

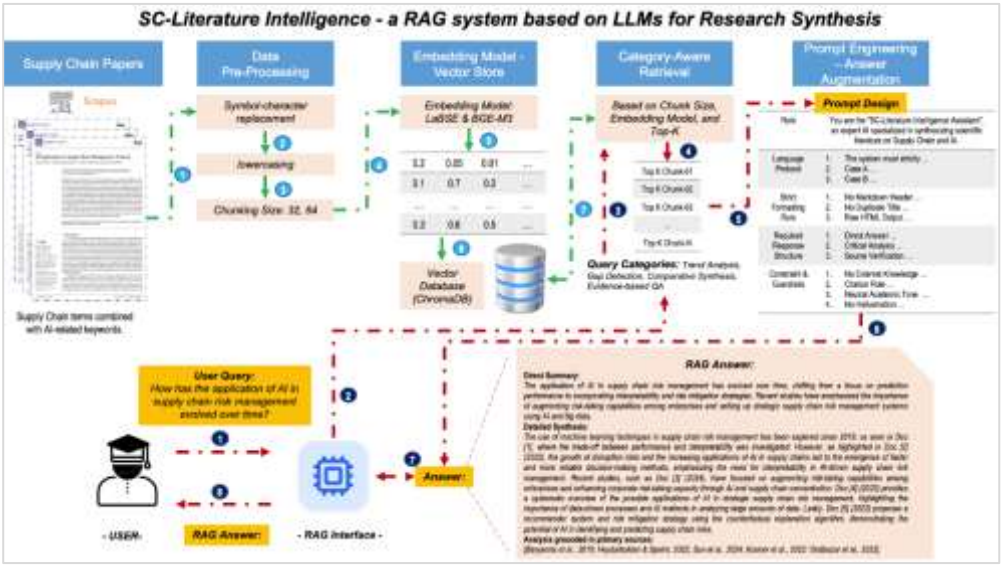

Figure 1. The proposed RAG architecture for research synthesis.

Data Acquisition and Data Pre-processing

The research data used in this paper, related to AI applications in the supply chain field, is sourced from Scopus. The search query used the term "supply chain" combined with AI-related terms, including artificial intelligence, AI, machine learning, deep learning, predictive analytics, big data, automation, smart warehouse, and digital twin. This process resulted in 6,683 publications retrieved from 2015 to 2025. Figure 2 presents the distribution and significant growth in the number of publications over the past 10 years. This indicates an increasing adoption of AI in the supply chain field. The corpus was deliberately restricted to abstracts, because abstracts are consistently available and they provide standardized, information-dense summaries of each study's objectives, methods, and findings. The trade-off of this is acknowledged as a limitation from the outset and is revisited in the Limitation section.

Data pre-processing prepares abstract data for use in building a RAG system. This stage consists of text cleaning to remove HTML tags, punctuation, special characters, character normalization, case folding to convert sentences to lowercase versions for consistency, and chunking to break text into smaller, manageable, and meaningful segments of 32 and 64 tokens (with a 10% overlap to maintain continuity). The chunk sizes of 32 and 64 tokens were selected to match the length characteristics of the corpus: because each document is an abstract of limited length, these segments preserve fine-grained retrieval granularity while still yielding several semantically coherent chunks per abstract. Substantially larger segments, such as 128 or 256 tokens, would approach or exceed the length of an entire abstract, collapsing each document into one or two chunks and effectively reducing chunk-level retrieval to document-level matching. An illustration of the data pre-processing process is presented in Table 2.

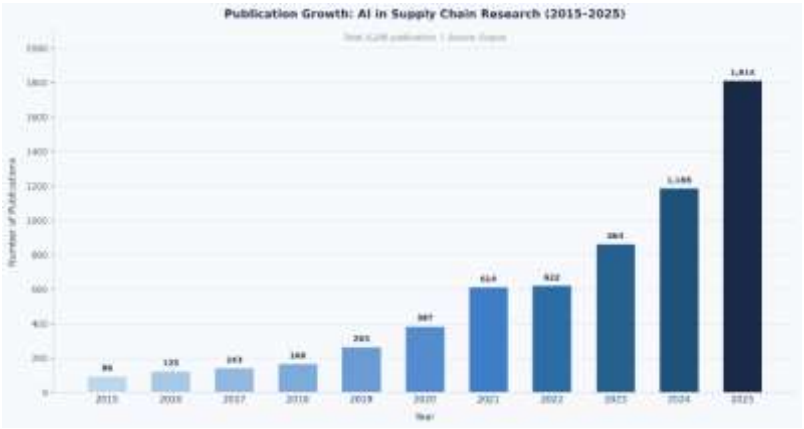

Figure 2. The Number of obtained publications per-year in the domain of AI application in supply chain.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16477
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 2. Data Pre-processing Illustration

Original Abstract "High-Entropy Alloys (HEAs), such as Al10.3Co17Cr7.5Fe9Ni48.6, have demonstrated superior performance compared to conventional materials—particularly in high-temperature engine applications. However, their development often relies on critical raw materials (CRMs), which pose sustainability & supply-chain risks." …
Text Cleaning "High Entropy Alloys HEAs such as Al Co Cr Fe Ni have demonstrated superior performance compared to conventional materials particularly in high temperature engine applications However their development often relies on critical raw materials CRMs which pose sustainability and supply chain risks" …
Case Folding "high entropy alloys heas such as al co cr fe ni have demonstrated superior performance compared to conventional materials particularly in high temperature engine applications however their development often relies on critical raw materials crms which pose sustainability and supply chain risks" …
Chunking Chunk 1: "high entropy alloys heas such as al co cr fe ni have demonstrated superior performance compared to conventional materials particularly in high temperature engine applications". Chunk 2: "however their development often relies on critical raw materials crms which pose sustainability and supply chain risks". …

Embedding Model

The embedding stage is the mapping of chunks to a high-dimensional vector space. Chunks with similar meanings will have close vector distances. This process allows for semantic search rather than just word-matching. The embedding model was built using Language-agnostic BERT Sentence Embedding (LaBSE) (Feng, Yang, Cer, Arivazhagan, & Wang, 2022) and BAAI General Embedding-Multi-Lingual, Multi-Functionality, and Multi-Granularity (BGE-M3) (Chen et al., 2024). LaBSE is a Transformer-based embedding model designed for multilingual sentence representation. This model is trained using the translation ranking objective, where texts that have the same meaning in different languages will have close vector distances. Although this model is stable for semantic similarity and cross-lingual retrieval, it is not designed to perform optimally in complex retrieval scenarios, such as multi-task or long-context reasoning. Similar to LaBSE, BGE-M3 is a Transformer-based model designed for various modern retrieval scenarios. This model has capabilities for dense retrieval, similarity learning, and multi-task learning. BGE-M3 is capable of capturing information at the word, sentence, and document levels. This model also supports multi-granularity and multi-lingual settings. BGE-M3 is optimized for high performance on retrieval tasks, including in long contexts and diverse domains. Furthermore, all embedding vectors based on both LaBSE and BGE-M3 are then stored in the ChromaDB vector database (Bashir et al., 2026).

Retrieval Process

The retrieval process begins with a user query in text form. This query is then converted into an embedding vector using the same model as the document vector in ChromaDB (LaBSE or BGE-M3), to ensure consistency in its representation. Next, the system calculates the similarity between the query embedding and all the chunk embeddings in the database using cosine similarity. Each chunk is then assigned a relevance score based on its similarity to the query. Based on this score, the system performs a ranking process and selects the Top-K chunks with the highest scores as the retrieval result. The K value is determined through experimentation. A small value can lead to incomplete information, while a large value can reduce the focus of the context. The selected chunks are then combined into a single context and used as additional input to the model. This context is then passed into the LLM prompt as part of the augmentation process. The retrieval information becomes the basis for the LLM to generate relevant, ground-truth-based answers.

Prompt Engineering

The retrieval context is integrated into the answer generation stage using prompt engineering and context injection techniques, which serve as an external knowledge base for the llama-3.1-8b-instant model [24]. This approach is strategically implemented to ensure the validity of evidence-based and consistent answers to the 6,683 scientific articles used, while mitigating the risk of information hallucinations that often occur in generative models without external context support. Table 3 shows the prompt engineering design designed to produce relevant and data-appropriate answers.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16477
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 3. Prompt Engineering Design.

01 — ROLE & OBJECTIVE	
Role	You are the "SC-Literature Intelligence Assistant", an expert AI specialized in synthesizing scientific literature on Supply Chain and AI.
Core Objective	Answer the user's question using ONLY the provided Context Data. Do not use outside knowledge under any circumstances.
02 — LANGUAGE PROTOCOL	
Language Protocol	The system must strictly follow the language of the User Query, regardless of the language used in the Context Data. • Case A — English Query: Output must be 100% English, with all findings translated from Indonesian context into English. • Case B — Indonesian Query: Output must be 100% Indonesian, with all findings translated from English context into Indonesian, while keeping technical terms in English (e.g., RAG, Bullwhip Effect, Robustness). • Never mix languages within a single response.
03 — STRICT FORMATTING RULES	
Output Format	• No Markdown: No Markdown headers (###, ##, or bolded titles). • No Duplicate Titles: Do not repeat section titles such as "Direct Answer" or "Direct Summary". • Raw HTML Only: The response must start directly with and contain valid HTML only. • Language Consistency: Never mix languages within a single output block.
04 — REQUIRED RESPONSE STRUCTURE	
① Direct Answer	A concise response (maximum 3 sentences) addressing the core question. HTML structure must follow a predefined format with emphasis on key concepts.
② Critical Analysis	A detailed synthesis presented using bullet points, explaining how and why findings emerge from the context, with every claim supported by inline numeric citations.
③ Source Verification	A closing note listing primary authors used in the analysis to ensure transparency and traceability of sources.
05 — CONSTRAINTS & GUARDRAILS	
Guardrails	• No Outside Knowledge: If information is not found in the provided context, the system must explicitly state that it is unavailable. • Citation Rule: Every claim in the Critical Analysis section must include a numeric citation. • Neutral Academic Tone: Responses must remain objective, analytical, and scholarly throughout. • No Hallucinations: The system must honestly acknowledge missing information rather than fabricate or infer answers.

Experiment Scenario

The RAG system evaluation was conducted using the RAGAS framework (Es, James, Espinosa Anke, & Schockaert, 2024). across four metrics: Faithfulness measures factual consistency by verifying each claim in the answer against retrieved chunks (score = supported claims ÷ total claims); Answer Relevancy measures semantic alignment between the answer and query intent via reverse question generation and cosine similarity; Context Precision measures retrieval noise by labeling each Top-K chunk as relevant or irrelevant (score = relevant chunks ÷ total retrieved chunks); and Context Recall measures concept coverage by checking whether the query's core concepts appear in the Top-K (score = covered concepts ÷ total core concepts).

The experiment was structured as follows. Twenty queries were constructed per category across four query types, namely trend analysis, gap detection, comparative synthesis, and evidence-based QA, yielding 80 queries in total. Each query was evaluated across all combinations of two embedding models (BGE-M3 and LaBSE), two chunk sizes (32 and 64 tokens), and two Top-K values (3 and 5), producing 8 configurations per query (2 × 2 × 2) and a total of 640 experimental runs. The Top-K values of 3 and 5 were selected to balance context sufficiency against context dilution: retrieving fewer than three chunks risks omitting complementary evidence, whereas progressively larger values introduce weakly related chunks that dilute Context Precision, a noise-sensitivity effect documented for RAG pipelines (Fan et al., 2024). Restricting the grid to 2 × 2 × 2 therefore concentrates the

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16477
e-ISSN : 2541-2019
p-ISSN : 2541-044X

evaluation on the region where retrieval quality is most sensitive while keeping the 640-run design computationally tractable. Finally, failure analysis was conducted on cases scoring below the acceptable threshold (RAGAS < 0.5), categorizing failures at two stages: retrieval (hard miss, partial miss, context noise) and generation (incomplete answer, abstractive deviation, semantic drift, hallucination). This threshold was adopted based on the proportional structure of RAGAS metrics, in which each metric is bounded within the range [0, 1] and a score below 0.5 indicates that the majority of evaluated quality dimensions fail to meet the minimum acceptable criterion. Table 4 presents sample queries per category.

Table 4. Sample Queries for Each Query Category.

No.	Query Category	Sample Queries
1	Trend Analysis (5 queries)	1. How has the application of AI in supply chain risk management evolved over time? 2. What were the dominant AI techniques used in early research on supply chain resilience? 3. What changes are observed in evaluation metrics across decades? 4. How has sustainability been incorporated into AI-based resilience models? 5. What emerging trends define the current era of AI for supply chain resilience?
2	Gap Detection (5 queries)	1. What aspects of supply chain risk are underexplored in current AI research? 2. Which AI techniques are rarely applied to supply chain resilience problems? 3. What types of supply chain disruptions are missing from existing studies? 4. Are there gaps in AI research for small and medium-sized supply chains? 5. What geographic regions are underrepresented in AI-based SCRM research?
3	Comparative Synthesis (5 queries)	1. How do machine learning and deep learning approaches differ in SCRM applications? 2. Compare AI models used for disruption prediction versus resilience optimization. 3. How do rule-based systems compare with data-driven AI in risk management? 4. What differences exist between supervised and unsupervised models in SCRM? 5. How do predictive and prescriptive AI models differ in objectives?
4	Evidence- based QA (5 queries)	1. What datasets are commonly used in AI-based SCRM studies? 2. What performance metrics are most frequently reported? 3. What empirical improvements does AI bring to supply chain resilience? 4. What reduction in disruption impact is reported using AI models? 5. What sample sizes are used in empirical SCRM studies?

RESULT

The Output of SC-Literature Intelligence

This section presents an illustrative output from the best-performing configuration (BGE-M3, Chunk 64, Top-K 5) to demonstrate the system's end-to-end behaviour across retrieval, generation, and evaluation dimensions, as shown in Table 5.

Table 5. Sample Query Design.

Query Category: Trend Analysis | Configuration: BGE-M3, Chunk 64, Top-K 5	
① User Query	
How has the application of AI in supply chain risk management evolved over time?	
② RAG System Answer	
Direct Summary	The application of AI in supply chain risk management has evolved over time, shifting from a focus on prediction performance to incorporating interpretability and risk mitigation strategies. Recent studies have emphasized the importance of augmenting risk-taking capabilities among enterprises and

*name of corresponding author

	setting up strategic supply chain risk management systems using AI and big data.
Detailed Synthesis	The use of machine learning techniques in supply chain risk management has been explored since 2019, as seen in Doc [1], where the trade-off between performance and interpretability was investigated. As highlighted in Doc [2] (2022), the growth of disruption risks and increasing applications of AI led to the emergence of faster and more reliable decision-making methods, emphasizing the need for interpretability in AI-driven risk management. Recent studies, such as Doc [3] (2024), focus on augmenting risk-taking capabilities among enterprises and enhancing corporate risk-taking capacity through AI and supply chain concentration. Doc [4] (2022) provides a systematic overview of AI applications in strategic supply chain risk management, highlighting the importance of data-driven processes for analyzing large amounts of data. Doc [5] (2022) proposes a recommender system and risk mitigation strategy using the counterfactual explanation algorithm, demonstrating AI's potential in identifying and predicting supply chain risks.
Primary Sources	Baryannis et al., 2019; Heydarbakian & Spehri, 2022; Sun et al., 2024; Kramer et al., 2022; Ordibazar et al., 2022

③ Top-K Retrieved Chunks — Top-K = 5, Cosine Similarity Score

Rank	Title & Authors	Retrieved Chunk (excerpt)
#1 0.7106	Predicting Supply Chain Risks Using Machine Learning Baryannis, G.; Dani, S.; Antoniou, G.	managing supply chain risks has received increased attention in recent years aiming to shield supply chains from disruptions by predicting their occurrence and mitigating their adverse effects at the same time the resurgence of artificial intelligence ai has led to the investigation of machine learn...
#2 0.6964	Interpretable Machine Learning to Improve Supply Chain Resilience: An Industry 4.0 Recipe Heydarbakian, S.; Spehri, M.	the growth of disruption risks that is risks with very low probability of occurrence and very high adverse impacts eg pandemics earthquakes etc across the global supply chains has increased over the past decades in industry 40 environment the increasing applications of artificial intelligence ai thr...
#3 0.6811	Uncovering the Interactions Between Enterprise AI Transformation, Supply Chain Concentration and Corporate Risk-Taking Capacity Sun, Y.; Wan, L.; Kumar Mangla, S.K.; Xu, X.; Song, M.	amidst recent supply chain disruptions triggered by pandemics and crises the imperative of bolstering supply chain security and resilience is escalating central to this endeavor is the augmentation of risktaking capabilities among enterprises at supply chain nodes rapid advances in artificial intell...
#4 0.6794	Strategic Supply Chain Risk Management: Artificial Intelligence and Big Data Kramer, K.J.; Mousavi, D.; Schmidt, M.	complexity and uncertainty along supply chains can be made more manageable through datadriven processes artificial intelligence ai methods in particular can be used to analyse large amounts of data from companies as a result a strategic supply chain risk management can be set up to monitor various s...
#5 0.6717	A Recommender System and Risk Mitigation Strategy for SCM Using the Counterfactual	supply chain management scm and its disruptions and risks have been the focus of many researchers in recent times it is important how to identify these disruptions and risks to avoid them therefore many risk mitigation strategies

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16477
e-ISSN : 2541-2019
p-ISSN : 2541-044X

	Explanation Algorithm Ordibazar, A.H.; Hussain, O.; Saberi, M.	have been developed artificial intelligence ai is a powerful tool to i...	

④ RAGAS Evaluation Scores — BGE-M3 · Chunk 64 · Top-K 5

1.000 100% Faithfulness	0.980 98% Answer Relevancy	1.000 100% Context Precision	1.000 100% Context Recall

★★Perfect retrieval: Faithfulness, Context Precision, and Context Recall all achieved maximum scores (1.000), indicating that the retrieved chunks were perfectly relevant, complete, and factually grounded. Answer Relevancy (0.980) confirms near-perfect semantic alignment with the query intent.

RAGAS Evaluation Result Focused on Embedding Models

Figure 3 presents the results of experiments testing the system performance with variations in Top-K parameters, chunk size, and the same LLM model (llama-3.1-8b-instant) focusing on embedding models. In general, BGE-M3 shows stronger and more stable performance than LaBSE, for all RAGAS indicators. Increasing the Top-K value from 3 to 5 relatively improves the system performance on both LaBSE and BGE-M. The BGE-M3 model is relatively more able to utilize additional context more effectively. The 64 chunk and 5 Top-K configuration on BGE-M3 produces the best overall performance. This advantage is explained by the training design of the two models. BGE-M3 is optimized directly for retrieval through multi-granularity, multi-functionality dense-embedding objectives with self-knowledge distillation (Chen et al., 2024), producing embeddings that discriminate the technical, domain-specific vocabulary of supply chain AI literature. LaBSE, in contrast, is trained with a translation-ranking objective aimed at cross-lingual sentence alignment (Feng, Yang, Cer, Arivazhagan, & Wang, 2022), which favours cross-language equivalence over within-domain semantic discrimination. The direction of the performance gap is consistent across all four RAGAS metrics and all eight configurations, which strengthens the reliability of this observation; a formal inferential test was not applied and this is acknowledged in the Limitation section.

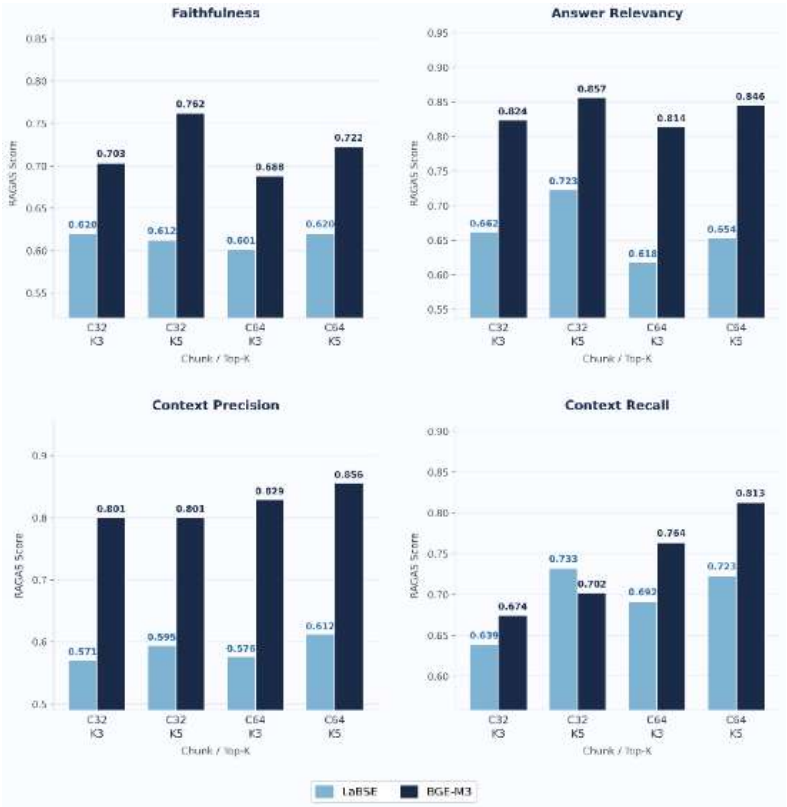

Figure 3. Comparison of RAGAS evaluation scores (Faithfulness, Answer Relevancy, Context Precision, and Context Recall) per Configuration by Query Category.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16477
e-ISSN : 2541-2019
p-ISSN : 2541-044X

RAGAS Evaluation Focused on Query Category

Figure 4 presents the results of experiments focusing on embedding models. In general, increasing Top-K improves context recall, especially in Trend Analysis and Gap Detection. Furthermore, the combination of higher chunks and Top-K relatively improves performance on queries requiring synthesis and proof. Gap Detection queries show consistent performance in context precision, indicating that the retrieved context tends to be relevant for the Top-K variation. Comparative Synthesis has high answer relevance, but is often accompanied by lower context recall. In the Evidence-based QA category, a chunk size of 64 with Top-K of 5 shows a balance between faithfulness and context recal.

DISCUSSION

Overall System Viability

Across 640 experimental queries, the system passed all four RAGAS thresholds in 292 cases (45.62%), while the remaining 348 queries (54.38%) exhibited at least one detectable failure. This indicates that the system operates at a functional level for research synthesis, with performance strongly dependent on the choice of embedding model. BGE-M3 achieved a pass rate of 58.75% (188 of 320), compared to 32.50% for LaBSE (104 of 320), a gap of 26.25 percentage points. The best configuration, BGE-M3 with chunk size 64 and Top-K 5, reached the highest pass rate at 62.50% (50 of 80), confirming that larger context windows and broader retrieval jointly improve reliability. Table 6 summarizes these results by embedding model, including pass rates and failure rates at both the retrieval and generation stages. Overall, the system is operationally viable for structured research synthesis in the supply chain AI domain, particularly when paired with BGE-M3.

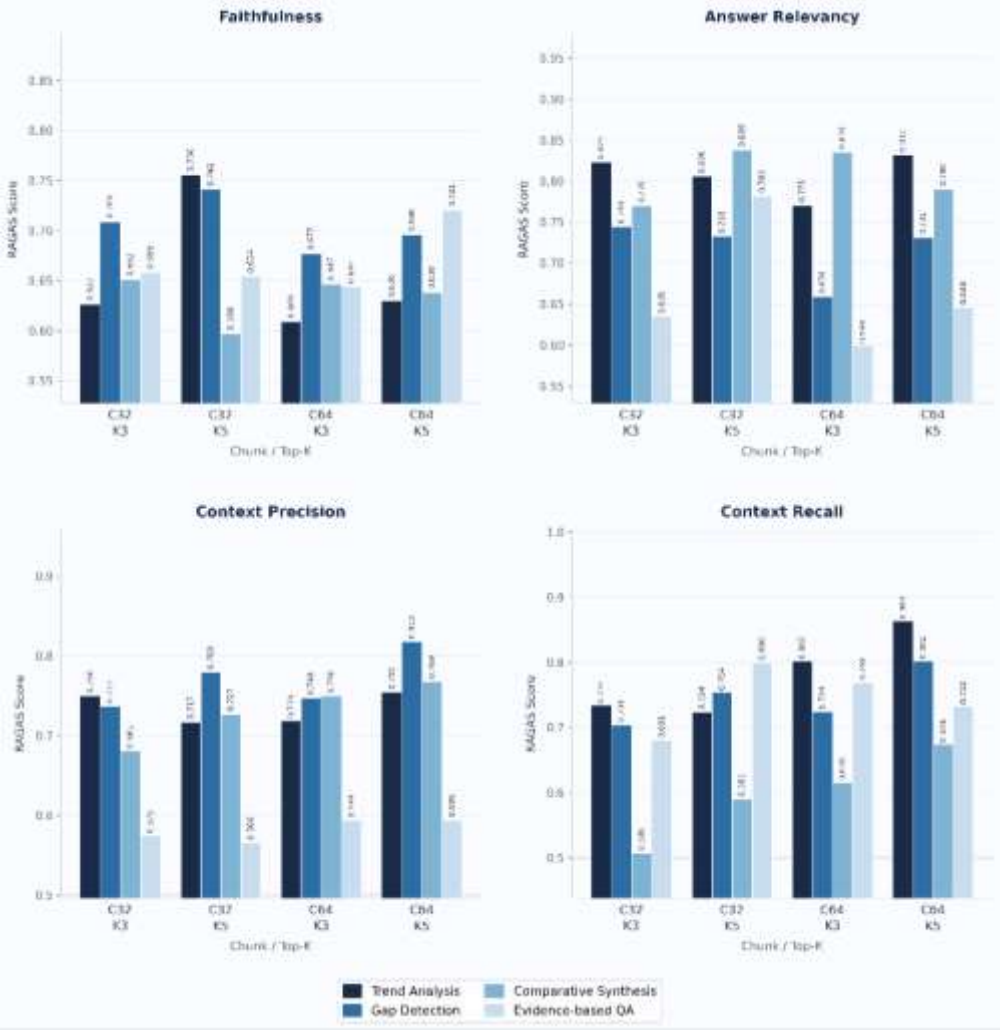

Figure 4. Comparison of RAGAS evaluation scores (Faithfulness, Answer Relevancy, Context Precision, and Context Recall) across embedding model configurations with varying chunk sizes and Top-K retrieval values.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16477
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 6. Stage-Level Comparison of LaBSE and BGE-M3 (n = 320 per model).

Metric	LaBSE (n=320)	BGE-M3 (n=320)	Difference (BGE-M3 − LaBSE)	Advantage
Pass Rate	32.50%	**58.75%**	+26.25 pp	BGE-M3
Retrieval-Stage Failure	52.50%	**30.31%**	−22.19 pp	BGE-M3
Generation-Stage Failure	15.00%	**10.94%**	−4.06 pp	BGE-M3
Retrieval Failure	26.88%	**22.19%**	−4.69 pp	BGE-M3
Context Noise	25.62%	**8.12%**	−17.50 pp	BGE-M3
Hallucination	12.81%	**10.00%**	−2.81 pp	BGE-M3
Semantic Drift	2.19%	**0.94%**	−1.25 pp	BGE-M3

Failure Stage Analysis: Retrieval vs Generation

Retrieval-stage failures comprising Retrieval Failure (Context Recall < 0.5) and Context Noise (Context Recall ≥ 0.5 but Context Precision < 0.5) account for the dominant share of system errors, totaling 265 out of 348 failed queries (76.15% of all failures, or 41.41% of all 640 queries). Retrieval Failure alone affected 157 queries (24.53%), with a mean Context Recall of 0.173, indicating that the system retrieved chunks covering fewer than one-fifth of the core query concepts on average in these cases. Context Noise affected 108 queries (16.88%), characterized by adequate recall (mean = 0.847) but severely degraded precision (mean = 0.134), meaning the retrieved set contained relevant chunks but was heavily diluted by irrelevant content. The contrast between LaBSE and BGE-M3 is most pronounced at the retrieval stage. LaBSE exhibited retrieval-stage failures in 168 out of 320 queries (52.50%), while BGE-M3 experienced the same in only 97 queries (30.31%). This 22-point gap strongly suggests that BGE-M3's multi-granularity, multi-task training enables more accurate semantic matching in the supply chain AI domain, where terminology is technical and domain-specific. LaBSE, while effective for cross-lingual tasks, produces embeddings that are less discriminative for within-domain precision retrieval in this context.

Generation-stage failures comprising Hallucination (Faithfulness < 0.5) and Semantic Drift (Answer Relevancy < 0.5 with all other thresholds met) affected 83 queries in total (12.97% of all 640 queries). Hallucination was the dominant generation failure with 73 cases (11.41%), while Semantic Drift was comparatively rare at 10 cases (1.56%). Notably, queries classified as Hallucination exhibited a mean Faithfulness of 0.267 indicating that nearly three-quarters of the claims in these answers were unsupported by the retrieved context while their mean Context Precision (0.904) and Context Recall (0.909) were near-perfect. This pattern confirms that hallucination in this system is primarily a generation-stage phenomenon: the LLM produces unfaithful answers even when the retrieval stage performs correctly, suggesting that prompt guardrails rather than retrieval improvements are the appropriate intervention.Semantic Drift cases (n=10) exhibited the expected profile: mean Faithfulness of 0.794 and mean Context Precision of 0.817 confirm factual grounding, while mean Answer Relevancy of 0.276 reveals that the generated answers deviated substantially from the core query intent. The rarity of this failure mode (1.56%) suggests that the prompt design effectively constrains response focus in most cases.

As summarized in Table 7, the analysis reveals a clear asymmetry between the two failure stages. Retrieval-stage failures (41.41%) are more than three times as frequent as generation-stage failures (12.97%). This points to a direct design implication. Improving embedding quality is likely to yield greater gains than refining prompt engineering. The most impactful intervention would be fine-tuning the embedding model on a supply chain domain corpus to reduce the high retrieval failure rate, particularly for LaBSE. Hallucination, by contrast, occurs mainly when retrieval succeeds. This suggests that the current prompt guardrails, such as the explicit no-outside-knowledge constraint and citation requirements, are partially effective but not enough to fully prevent parametric knowledge leakage in the LLM.

Table 7. Failure Distribution by Stage and Type (n=640)

Stage	Failure Type	Queries (n)	% of 640
Retrieval	Retrieval Failure	157	24.53%
	Context Noise	108	16.88%
Subtotal		265	41.41%
Generation	Hallucination	73	11.41%
	Semantic Drift	10	1.56%
Subtotal		83	12.97%
Total		348	54.38%

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16477
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Query Category Sensitivity Analysis

Table 8 present the classification results disaggregated by query category, revealing substantial variation in system performance that cannot be explained by retrieval configuration alone. Pass rates range from 55.00% for Gap Detection to 36.25% for Evidence-based QA, a gap of 18.75 percentage points larger than the effect of any single configuration parameter. This confirms that query category is a primary determinant of system performance and warrants category-specific analysis.

Table 8. Classification Results by Query Category (n = 160 per category).

Query Category	Pass Rate	Retrieval-Stage Failure	Generation-Stage Failure	Dominant Failure Type
Gap Detection	55.00%	33.12%	11.88%	Retrieval Failure
Trend Analysis	48.12%	35.00%	16.88%	Retrieval Failure + Hallucination
Comparative Synthesis	43.12%	44.38%	12.50%	Retrieval Failure (38.75%)
Evidence-based QA	36.25%	53.12%	10.62%	Context Noise (32.50%)

Gap Detection achieved the highest pass rate (55% with 88 out of 160) and the lowest generation-stage failure rate (11.88%). BGE-M3 clearly outperformed LaBSE here (67.50% vs 42.50% pass) with a high mean Context Precision (0.850), reflecting that gap-oriented queries activate a narrow specific set of chunks that BGE-M3 retrieves precisely. This makes Gap Detection the query type best suited to the current architecture.

Trend Analysis reached a 48.12% pass rate (77 out of 160) with 35.00% retrieval-stage and 16.88% generation-stage failures. The 24 Hallucination cases (15%) are notable given a corpus spanning 2015–2025. These hallucinations likely come from queries about recent trends after 2024 that are still underrepresented in the corpus. BGE-M3 again led on Faithfulness (0.684 vs 0.627) and Context Precision (0.896 vs 0.574).

Comparative Synthesis exhibited the highest retrieval-stage failure rate at 44.38% (71 out of 160) driven predominantly by Retrieval Failure (38.75%, n=62). This pattern is consistent with the structural challenge inherent to comparative queries. They require the simultaneous retrieval of chunks representing multiple distinct entities or approaches,which is difficult to achieve at the abstract level. When a single query embedding must represent two or more contrasting concepts, cosine similarity tends to retrieve chunks that are generically relevant to the domain rather than specifically relevant to each comparative dimension. BGE-M3 partially mitigates this effect its Retrieval Failure rate for Comparative Synthesis (35.00%) is lower than LaBSE's (42.50%), but the structural limitation persists for both models. This finding directly supports the recommendation for full-text retrieval beyond abstract-level chunks for this query category.

Evidence-based QA had the lowest pass rate (36.25%, 58 out of 160), with retrieval-stage failures dominant (53.12%, 85 out of 160). Context Noise was the largest single failure mode (32.50%, 52 cases), concentrated in LaBSE (36 vs 16 for BGE-M3). For precise factual queries, LaBSE retrieves chunks with superficially relevant vocabulary but lacking the specific content needed. This suggests short abstract-level chunks are structurally inadequate for fact-retrieval requiring numerical or methodological detail from full-text sections.

These findings align with, and extend, prior evidence. The dominance of retrieval-stage failures corroborates the foundational premise of RAG that generation quality is bounded by retrieval quality (Lewis et al., 2020), and mirrors the retrieval noise sensitivity highlighted in the survey of Fan et al. (2024). The consistent superiority of BGE-M3 over LaBSE agrees with the multi-granularity retrieval optimization reported by Chen et al. (2024), while the weaker within-domain discrimination of LaBSE is consistent with its cross-lingual alignment objective (Feng et al., 2022). At the evaluation level, the failure patterns surfaced by the four metrics follow the metric semantics defined for RAGAS by Es et al. (2024): low Context Recall flags concept-coverage misses, whereas high recall combined with low precision flags context dilution, supporting the validity of the two-stage failure taxonomy applied here. The category-level variation observed in this study, however has no direct counterpart in these prior works, which evaluate RAG on conventional QA; this contrast reinforces the central claim that category-aware evaluation reveals failure structure that single-task benchmarks cannot expose.

Implication for RAG System Design

Two design implications emerge from this analysis. First, embedding model selection is the single most impactful design decision for this system. The 26-point pass rate gap between BGE-M3 and LaBSE driven almost entirely by retrieval-stage performance indicates that semantic representation quality determines system ceiling more than any other component. Future iterations should explore domain-adapted embedding models trained or

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

fine-tuned on supply chain and operations management literature to further reduce retrieval failure rates, particularly for Comparative Synthesis and Evidence-based QA.

Second, granularity of retrieved content must be matched to query type. The consistently high retrieval failure rates for Comparative Synthesis and Evidence-based QA suggest that 64-token abstract chunks are insufficient for multi-entity comparison and fact retrieval. A hybrid retrieval strategy using abstract-level chunks for Trend Analysis and Gap Detection, and full-text section-level chunks for Comparative Synthesis and Evidence-based QA is a promising direction that this study's findings directly motivate.

Limitation

Several limitations should be acknowledged. First, the classification framework applied in this analysis (Rules 1–4) relies entirely on RAGAS thresholds and does not incorporate human evaluation. The boundary between queries that pass all thresholds and those that are genuinely fully correct cannot be determined from RAGAS scores alone, and future work should incorporate structured annotation to validate this distinction. Second, the corpus consists exclusively of abstracts, which constrains the system's ability to retrieve specific factual details required for Evidence-based QA, a limitation that manifests directly in the high Context Noise rate for this category. Third, the findings are specific to the supply chain AI domain and may not generalize to other research domains with different vocabulary distributions or corpus characteristics. Fourth, the comparative claims in this study rest on descriptive aggregate scores across the 640 runs, inferential statistical validation such as paired t-tests or Wilcoxon signed-rank tests on per-query score distributions, was not performed and should be applied in future work to formally establish the significance of the observed differences between models and configurations.

CONCLUSION

This paper has developed a RAG system for research synthesis, focusing on scientific papers related to AI in the supply chain domain, motivated by the fragmented and voluminous nature of current scientific literature. The system supports four query categories, namely trend analysis, gap detection, comparative synthesis, and evidence-based QA, evaluated through 640 experimental runs using the RAGAS framework. The results demonstrate that BGE-M3 with chunk size 64 and Top-K 5 delivers the strongest overall performance, achieving Faithfulness of 0.722, Answer Relevancy of 0.845, Context Precision of 0.856, and Context Recall of 0.813. Among query categories, gap detection yields the highest retrieval quality with a peak Context Precision of 0.914, while comparative synthesis remains the most challenging category with Context Recall as low as 0.507. Failure analysis across 640 queries confirms that retrieval-stage issues, particularly retrieval failure (24.53%) and context noise (16.88%), dominate over generation-stage failures (12.97%), establishing embedding quality and domain-specific semantic matching as the primary bottlenecks for future improvement.

For future research, several concrete directions are recommended: (1) integrating full-text retrieval beyond abstract-level chunks to improve context completeness for comparative synthesis queries; (2) exploring hybrid retrieval combining dense and sparse methods such as BM25 and BGE-M3 to address the structural recall limitations observed in multi-entity queries; (3) fine-tuning the embedding model on supply chain domain corpora to reduce the semantic gap between query vocabulary and indexed content; and (4) conducting systematic prompt sensitivity analysis to quantify how prompt variation affects faithfulness and answer completeness across different query categories. These directions aim to address the identified failure patterns while extending the system's generalizability beyond the supply chain domain. Beyond these directions, the principal contribution of this work to the community is a category-aware research synthesis framework, coupled with a two-stage failure taxonomy that reframes RAG evaluation from conventional single-task question answering toward structured, multi-category literature intelligence providing a reusable evaluation design and domain-level evidence for embedding model selection in AI-driven supply chain research.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support and funding provided by the Internal Research Grant of Universitas Muhammadiyah Malang (UMM) 2026. The authors also thank UMM for the resources and support that made this research possible.

REFERENCES

Abogunrin, S., Slob, B. P. H., Lane, M., Emamipour, S., Twardowski, P., Boersma, C., & van der Schans, J. (2025). The introduction and adoption of artificial intelligence in systematic literature reviews: a discrete choice experiment. *BMJ Open*, *15*(10). https://doi.org/10.1136/bmjopen-2025-099921

Amugongo, L. M., Mascheroni, P., Brooks, S., Doering, S., & Seidel, J. (2025). Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, *4*(6 JUNE). https://doi.org/10.1371/journal.pdig.0000877

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Aytar, A. Y., Kilic, K., & Kaya, K. (2024). *A Retrieval-Augmented Generation Framework for Academic Literature Navigation in Data Science*. Retrieved from http://arxiv.org/abs/2412.15404

Bashir, H., Escriva, R., Isom, M., Huber, J., Macronova, Kedia, S., … naynaly10. (2026, March). *chroma-core/chroma*. Retrieved from https://github.com/chroma-core/chroma

Besrour, I., He, J., Schreieder, T., & Färber, M. (2025). SQuAI: Scientific Question-Answering with Multi-Agent Retrieval-Augmented Generation. *CIKM 2025 - Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 6603–6608. Association for Computing Machinery, Inc. https://doi.org/10.1145/3746252.3761471

Bolaños, F., Salatino, A., Osborne, F., & Motta, E. (2024). Artificial intelligence for literature reviews: opportunities and challenges. *Artificial Intelligence Review*, *57*(9). https://doi.org/10.1007/s10462-024-10902-3

Che, T.-Y., Mao, X.-L., Lan, T., & Huang, H. (2024). A Hierarchical Context Augmentation Method to Improve Retrieval-Augmented LLMs on Scientific Papers. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 243–254. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3637528.3671847

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 2318–2335). Bangkok, Thailand: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.137

Daios, A., Kladovasilakis, N., Kelemis, A., & Kostavelis, I. (2025). AI Applications in Supply Chain Management: A Survey. *Applied Sciences (Switzerland)*, *15*(5). https://doi.org/10.3390/app15052775

Dindarian, A. (2026). Harnessing AI and Big Data to Build a Resilient Supply Chain: An Overview. *International Series in Operations Research and Management Science*, *373*, 233 – 251. https://doi.org/10.1007/978-3-032-01218-0_9

Douaioui, K., Oucheikh, R., Benmoussa, O., & Mabrouki, C. (2024). Machine Learning and Deep Learning Models for Demand Forecasting in Supply Chain Management: A Critical Review. *Applied System Innovation*, *7*(5). https://doi.org/10.3390/asi7050093

Ehrenthal Joachim C. F. and Gachnang, P. and L. L. and R. H. and S. F. (2024). Integrating Generative Artificial Intelligence into Supply Chain Management Education Using the SCOR Model. In C. and K. C. Almeida João Paulo A. and Di Ciccio (Ed.), *Advanced Information Systems Engineering Workshops* (pp. 59–71). Cham: Springer Nature Switzerland.

Elliott, J., & Turner, T. (2022). Innovations in Systematic Review Production. In *Systematic Reviews in Health Research: Meta-Analysis in Context: Third Edition*. https://doi.org/10.1002/9781119099369.ch23

Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (2024). RAGAs: Automated Evaluation of Retrieval Augmented Generation. In N. Aletras & O. De Clercq (Eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 150–158). St. Julians, Malta: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-demo.16

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., … Li, Q. (2024). A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6491–6501. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3637528.3671470

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT Sentence Embedding. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 878–891). Dublin, Ireland: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.62

Galán-Mena, J., López-Nores, M., Galán, J., Pulla-Sánchez, D., Guerrero-Vásquez, L. F., & Salgado-Guerrero, J. P. (2026). Topic Analysis in News About AI: An Approach Based on LDA and Large Language Models. *Lecture Notes in Networks and Systems*, *1582 LNNS*, 355 – 366. https://doi.org/10.1007/978-3-032-01130-5_28

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., … Wang, H. (2024). *Retrieval-Augmented Generation for Large Language Models: A Survey*. Retrieved from http://arxiv.org/abs/2312.10997

George, L., & Sumathy, P. (2023). An integrated clustering and BERT framework for improved topic modeling. *International Journal of Information Technology (Singapore)*, *15*(4), 2187 – 2195. https://doi.org/10.1007/s41870-023-01268-w

Godinez, A. (2025). *HYSEMRAG: A HYBRID SEMANTIC RETRIEVAL-AUGMENTED GENERATION FRAMEWORK FOR AUTOMATED LITERATURE SYNTHESIS AND METHODOLOGICAL GAP ANALYSIS A PREPRINT*. Retrieved from https://arxiv.org/abs/2508.05666

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Gokdemir, O., Siebenschuh, C., Brace, A., Wells, A., Hsu, B., Hippe, K., … Ramanathan, A. (2025). HiPerRAG: High-Performance Retrieval Augmented Generation for Scientific Insights. *Proceedings of the Platform for Advanced Scientific Computing Conference*, 1–13. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3732775.3733586

Hou Yu and Bishop, J. R. and L. H. and Z. R. (2025). Improving Dietary Supplement Information Retrieval: Development of a Retrieval-Augmented Generation System With Large Language Models. *J Med Internet Res*, *27*, e67677. https://doi.org/10.2196/67677

Hoyle, A., Goel, P., Peskov, D., Hian-Cheong, A., Boyd-Graber, J., & Resnik, P. (2021). Is automated topic model evaluation broken? the incoherence of coherence. *Proceedings of the 35th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.

Hu, Y., Lei, Z., Dai, Z., Zhang, A., Angirekula, A., Zhang, Z., & Zhao, L. (2025). CG-RAG: Research Question Answering by Citation Graph Retrieval-Augmented LLMs. *SIGIR 2025 - Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 678–687. Association for Computing Machinery, Inc. https://doi.org/10.1145/3726302.3729920

Huang, K., Chen, B., Lu, Y., Wu, S., Wang, D., Huang, Y., … Peng, X. (2025). Lifting the Veil on Composition, Risks, and Mitigations of the Large Language Model Supply Chain. *Arxiv*. Retrieved from https://arxiv.org/abs/2410.21218

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., … Liu, T. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, *43*(2). https://doi.org/10.1145/3703155

Jackson, I., Saénz, M. J., Ivanov, D., & Ma, B. J. (2026). Supply chain mapping through retrieval-augmented generation: applications to the electronics industry. *Journal of the Operational Research Society*, *0*(0), 1–21. https://doi.org/10.1080/01605682.2025.2608868

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., … Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, *55*(12). https://doi.org/10.1145/3571730

Krotkov, N. A., Sbytov, D. A., Chakhoyan, A. A., Kornienko, P. I., Starikova, A. A., Stepanov, M. G., … Skorb, E. V. (2025). Nanostructured Material Design via a Retrieval-Augmented Generation (RAG) Approach: Bridging Laboratory Practice and Scientific Literature. *Journal of Chemical Information and Modeling*, *65*(20), 11064–11078. https://doi.org/10.1021/acs.jcim.5c01897

Lee, D., Sohn, S. S., Lee, B., & Kim, D. (2025). Domain-aligned LLM framework for trustworthy scientific Q/A via query reformulation retrieval-augmented generation. *ChemRxiv*, *2025*(1127). https://doi.org/10.26434/chemrxiv-2025-t5dqd

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., … Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.

Li, X., Shang, W., Niu, Z., Xie, B., & Zhu, L. (2025). Can large language models evaluate suppliers' qualifications of complex production lines? An approach incorporating retrieval-augmented generation and Chain-of-Thought. In T. Stafford, R. A. Syler, B. C. Y. Tan, M. K. Ahuja, C. Hsu, & E. A. Whitley (Eds.), *Proceedings of the 46th International Conference on Information Systems, ICIS 2025, Achieving Digital Integration in the Age of AI, Nashville, TN, USA, December 14-17, 2025*. Association for Information Systems. Retrieved from https://aisel.aisnet.org/icis2025/da_bus/da_bus/5

Li, Y., Ko, H., & Ameri, F. (2025). Integrating Graph Retrieval-Augmented Generation With Large Language Models for Supplier Discovery. *Journal of Computing and Information Science in Engineering*, *25*(2). https://doi.org/10.1115/1.4067389

Neha, F., Bhati, D., & Shukla, D. K. (2025). Retrieval-Augmented Generation (RAG) in Healthcare: A Comprehensive Review. *AI*, *6*(9). https://doi.org/10.3390/ai6090226

Oliva, M. P. (2024). *Evaluation Metrics for Retrieval Augmented Generation in the Scientific Domain*. Retrieved from https://addi.ehu.es/handle/10810/73106

Ponnock, J., Kenneally, G., Briggs, M. R., Yeo, E., III, T. P., Kinberg, N., … Hechtman, D. (2025). Real-Time RAG for the Identification of Supply Chain Vulnerabilities. *Arxiv*. Retrieved from https://arxiv.org/abs/2509.10469

Riad, M., Naimi, M., & Okar, C. (2024). Enhancing Supply Chain Resilience Through Artificial Intelligence: Developing a Comprehensive Conceptual Framework for AI Implementation and Supply Chain Optimization. *Logistics*, *8*(4). https://doi.org/10.3390/logistics8040111

Saba, W., Wendelken, S., & Shanahan, J. (n.d.). *Question-Answering Based Summarization of Electronic Health Records using Retrieval Augmented Generation*. Retrieved from https://www.trychroma.com/

Sajjadi Mohammadabadi, S. M., Kara, B. C., Eyupoglu, C., Uzay, C., Tosun, M. S., & Karakuş, O. (2025). A Survey of Large Language Models: Evolution, Architectures, Adaptation, Benchmarking, Applications, Challenges, and Societal Implications. *Electronics*, *14*(18). https://doi.org/10.3390/electronics14183580

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Shamsuddoha, M., Khan, E. A., Chowdhury, M. M. H., & Nasir, T. (2025). Revolutionizing Supply Chains: Unleashing the Power of AI-Driven Intelligent Automation and Real-Time Information Flow. *Information (Switzerland)*, *16*(1). https://doi.org/10.3390/info16010026

Shi, Y., Xu, S., Yang, T., Liu, Z., Liu, T., Li, X., & Liu, N. (n.d.). *MKRAG: Medical Knowledge Retrieval Augmented Generation for Medical Question Answering*.

Simchi-Levi, D., Mellou, K., Menache, I., & Pathuri, J. (2025). *Large Language Models for Supply Chain Decisions*. Retrieved from https://arxiv.org/abs/2507.21502

Storan Dylan and Chiarelli, A. (2026, January). *Safeguarding Scholarly Communication Publisher Practices to Uphold Research Integrity*. International Association of Scientific, Technical and Medical Publishers. Retrieved from https://policycommons.net/artifacts/42179457/stm-research-integrity-storytelling_final-shared-6th-jan-2026/

Wagner, G., Lukyanenko, R., & Paré, G. (2022). Artificial intelligence and the conduct of literature reviews. *Journal of Information Technology*, *37*(2), 209–226. https://doi.org/10.1177/02683962211048201

Xu, R., Liu, H., Nag, S., Dai, Z., Xie, Y., Tang, X., … He, Q. (2025). SimRAG: Self-Improving Retrieval-Augmented Generation for Adapting Large Language Models to Specialized Domains. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 11534–11550). Albuquerque, New Mexico: Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.575

Zhang, G., Xu, Z., Jin, Q., Chen, F., Fang, Y., Liu, Y., … Peng, Y. (2025). Leveraging long context in retrieval augmented language models for medical question answering. *Npj Digital Medicine*, *8*(1), 239. https://doi.org/10.1038/s41746-025-01651-w

Zhang Shiran and Phan, E. and V. P. and P. Q. and S. S. (2026). Retrieval-Augmented Generation for Medical Question Answering on a Heart Failure Dataset: Performance Analysis. *JMIR Form Res*, *10*, e84932. https://doi.org/10.2196/84932

Zhu, B., & Vuppalapati, C. (2024). Enhancing Supply Chain Efficiency Through Retrieve-Augmented Generation Approach in Large Language Models. *2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService)*, 117–121. https://doi.org/10.1109/BigDataService62917.2024.00025

Zhu Beilei and Vuppalapati, C. (2025). Integrating Retrieval-Augmented Generation with Large Language Models for Supply Chain Strategy Optimization. In K. and D. L. Arabnia Hamid R. and Ferens (Ed.), *Applied Cognitive Computing and Artificial Intelligence* (pp. 475–486). Cham: Springer Nature Switzerland.

*name of corresponding author