ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep297

Retrieval-Augmented Large Language Models for Real-Time Supply Chain Disruption Intelligence and Decision Support

Sohail Sayed[1]; Nauman Sayed[2]

[1;2]California State University Fullerton

Publication Date: 2026/09/19

Abstract: Large language models offer unprecedented analytical capability, but their knowledge is frozen at the last training date — rendering them unusable for organizations whose mission depends on emerging, timely information [1]. This paper argues that retrieval-augmented generation is the missing mechanism that converts LLMs from static reasoners into real-time supply chain disruption intelligence systems: retrieval supplies the current evidence, grounding supplies the facts, and agentic orchestration supplies the decision loop. We propose RAGENT-SC, a retrieval-augmented agentic framework coupling (i) a continuously updated multi-format knowledge layer (news streams, contracts, supplier records, operational KPIs, and a supply network knowledge graph), (ii) hybrid vector–graph retrieval with sublinear scalability, (iii) grounded generation with provenance metadata, (iv) agentic orchestration of monitoring, analysis, planning, and audit agents, and (v) a governance layer with hallucination verification and human-in-the-loop escalation. The framework consolidates the reported evidence envelope: a seven-agent LLM-based framework detects, analyses, and responds to disruptions across multi-tier networks with F1 scores between 0.962 and 0.991, completing end-to-end analyses in a mean of 3.83 minutes at a cost of $0.0836 per disruption — a reduction of more than three orders of magnitude relative to multi-day analyst-driven assessments, validated on a real-world case study [2]; graph-grounded LLM reasoning improves ROUGE-1 F1 from 0.42 to 0.67 — a 59% improvement over the best comparative methods [3]; hybrid vector–graph retrieval scales sublinearly, doubling query time from 0.09 s to 0.20 s when the node count quadruples [4]; and knowledge-graph-augmented RAG outperforms plain RAG on supply chain question answering, with smaller open-weight models closing the gap with state-of-the-art frontier models [5]. The paper argues that retrieval quality — not model scale — is the binding constraint on real-time disruption intelligence, and that grounded, governed RAG is the difference between an LLM that recalls and a system that acts.

Keywords: *Retrieval-Augmented Generation, Large Language Models, Supply Chain Disruption Intelligence, Knowledge Graphs, Decision Support Systems, Real-Time Analytics, Agentic AI, Supply Chain Risk Management.*

How to Cite: Sohail Sayed; Nauman Sayed (2026) Retrieval-Augmented Large Language Models for Real-Time Supply Chain Disruption Intelligence and Decision Support. *International Journal of Innovative Science and Research Technology*, 11(9), 429-440. https://doi.org/10.38124/ijisrt/26sep297

I. INTRODUCTION

Supply chains operate on information that ages in minutes. A port closure, a supplier bankruptcy filing, a customs disruption, or an extreme-weather event loses its decision value the moment it is stale — yet most disruption-intelligence workflows still rely on manual analyst review of documents and news, with assessments that take days to produce [2]. Large language models have transformed what machines can do with text — summarization, extraction, reasoning, planning — but their parametric knowledge terminates at the model's training cutoff, making them structurally incapable of reasoning over events that occurred yesterday [1].

Retrieval-augmented generation resolves this tension. RAG combines a retrieval module that locates relevant information in an external knowledge base with a generation module that crafts contextualized responses, enabling contextualized analysis, predictive insight, and data-driven decision-making across supply chain operations [6]. Critically, RAG is the established mechanism for reducing hallucinations — grounding generation in retrieved evidence — across question answering, dialogue, and fact verification [7]. Yet its use in supply chain scenarios remains limited, chiefly because of the lack of high-quality, domain-specific knowledge bases [7]. This paper addresses exactly that gap: how to build the knowledge base, retrieve from it in real time, and orchestrate agents over it so that disruption warnings become decisions.

This paper is the seventh in a research program on resilient, AI-driven commerce. Paper 1 built a multi-signal early-warning framework; Paper 2 advanced to agentic AI for

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep297

detection, decision, and recovery; Paper 3 formalized graph-cascading risk; Paper 4 supplied a spatiotemporal weather front end; Paper 5 specialized the framework to food security; and Paper 6 connected the portfolio to platform-engineering via failure prediction and self-healing. This paper unifies the portfolio's LLM threads into a retrieval-first architecture: the evidence layer that makes every downstream agent — the early-warning engine, the cascade predictor, the recovery planner — factually grounded in what is happening now.

➢ *The Contributions of this Paper are:*

- A formalization of real-time disruption intelligence as a retrieval-grounded generation problem with explicit timeliness, faithfulness, and decision-utility constraints.
- RAGENT-SC, a retrieval-augmented agentic framework combining a continuously updated knowledge layer, hybrid vector–graph retrieval, grounded generation with provenance, agentic orchestration, and governance.
- A consolidated analysis of reported RAG/LLM performance in supply chains — detection F1, latency and cost, retrieval scaling, and grounding gains — with worked calculations and clearly flagged illustrative extensions (Sections V–VI).
- A discussion of the research frontier: retrieval quality as the binding constraint, knowledge-graph grounding, benchmark standardization, and governed deployment (Sections VII–VIII).

II. RELATED WORK

➢ *LLMs Meet Supply Chain Management*

LLMs have demonstrated complex reasoning and tool-based decision-making that motivates their application to real-world supply chain workflows — yet supply chains require reliable long-horizon, multi-step orchestration grounded in domain-specific procedures, which remains challenging for current models; the SupChain-Bench benchmark reveals substantial gaps in execution reliability across models, motivating SOP-free frameworks such as SupChain-ReAct that autonomously synthesize executable procedures for tool use [8]. Domain-specialized SCM LLMs built on RAG achieve expert-level competence, passing standardized supply chain management examinations and beer game tests, and reproduce classical insights while uncovering novel behaviors around the bullwhip effect [7]. Industry perspectives summarize the transformation: LLM-based technology automates data discovery, insight generation, and scenario analysis, "reducing the time to make decisions from days to minutes" [9]. The convergence of foundation models and multi-agent systems is consistently identified as enabling generalist agents with multi-faceted decision-making — yet it remains underexplored in supply chains relative to healthcare and finance [10].

➢ *Retrieval-Augmented Generation*

RAG's architecture is well established: a dual-component system in which a retrieval module pinpoints relevant information from a knowledge base and a generation module crafts contextualized responses, extended through iterative retrieval strategies and tailored chunk optimization for supply chain contexts [6]. In supply chain analysis specifically, integrating RAG preprocessing with web-scraping technologies reduces the latency of incorporating new information into an augmented LLM, enabling timely analysis of disruption signals; experimental evidence shows that fine-tuning the embedding retrieval model consistently yields the most significant performance gains — underscoring that retrieval quality, not model size, is the pivot — and that adaptive iterative retrieval, which adjusts retrieval depth based on context, further enhances performance on complex supply chain queries [1]. RAG is also the documented remedy for hallucination: retrieval-grounded generation reduces hallucination across question answering, dialogue, and fact verification [7].

➢ *Knowledge Graphs as the Retrieval Substrate*

Text-vector retrieval alone oversimplifies the relational structure of supply chains; networks and knowledge graphs are dual representations, and structural network science can drive retrieval — a graph traverser guided by network centrality scores extracts economically salient risk paths, while "context shells" embed raw figures in natural language so quantitative data is fully intelligible to the LLM, enabling concise, explainable, context-rich risk narratives in real time without fine-tuning or a dedicated graph database [11]. Ontology-driven graph construction integrated with RAG improves supplier discovery, answer quality, and visibility for small and medium manufacturers [12]. A graph-based LLM framework that constructs a graph knowledge base, creates prompts for graph retrieval, and guides generation demonstrates the approach on entity classification, link prediction, and reasoning, improving ROUGE-1 F1 from 0.42 to 0.67 — a 59% gain over the average of the best comparative methods [3]. Knowledge-graph-augmented RAG frameworks show that structured, factual context improves accuracy, relevance, and robustness of LLM responses, with smaller open-weight models achieving notable gains that reduce the performance gap with larger models [5]. Evaluations in decision support confirm the synergy: integrating RAG with knowledge graphs improves decision accuracy, reasoning transparency, and context relevance compared with either technology alone, particularly for cross-domain reasoning and ambiguous queries [13]. In supply chain visibility research, knowledge-graph extraction with LLMs supports tracing tier-2 and tier-3 sourcing — the mine-to-battery-to-vehicle dependency chain in electric vehicles [14] — complementing the generative relationship-prediction results on knowledge graphs in the portfolio [15].

➢ *Agentic Orchestration and Real-Time Intelligence*

Multi-agent LLM systems operationalize retrieval. A seven-agent framework powered by LLMs and deterministic tools autonomously monitors, analyses, and responds to disruptions across extended supply networks: agents jointly detect disruption signals from unstructured news, map them to multi-tier supplier networks, evaluate network-based exposure, and recommend mitigations such as alternative sourcing, achieving F1 scores between 0.962 and 0.991 with end-to-end analyses in a mean of 3.83 minutes at $0.0836 per disruption — more than three orders of magnitude faster than

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep297

multi-day analyst assessments — with operational applicability demonstrated on the 2022 Russia–Ukraine conflict [2]. For natural-hazard risk, automated systems retrieve news about supply nodes and extract risk features with four LLMs (Llama3.1-8B, Gemma3-12B, DeepSeek-R1-14B, Phi4-14B) under three prompting strategies, with Llama3.1-8B and Phi4-14B achieving the highest human-evaluation and BERTScore results and DeepSeek-R1-14B the lowest; hazards spanning wider regions degrade performance because location–risk links become complex [16]. Large-corpus monitoring is practical: a Risk Monitor Agent over an Indian news repository of ~3.88 million rows (2001–2023) uses keyword pre-filtering, spaCy NER, and LLM-assisted classification (Gemini 1.5-flash) to recognize primary risk classifications and most-impacted elements, validated on 2020–2023 data [17]. Graph-orchestrated dashboards coordinate procurement-contract analysis, SKU rationalization, and scenario planning through RAG with Llama Index and FAISS under LangGraph, with lightweight models synthesizing outputs into actionable summaries [18]. Consensus-seeking LLM agents reduce the bullwhip effect through conversation and, with tools, outperform restocking and centralized demand policies [19]; memory-retrieval agent architectures outperform benchmarks across supply chain scenarios [20]; and agentic decision prompts with episodic memory retrieval strengthen the LLM thread of the portfolio [21].

➢ *Llms in Decision Support and Operations Research*

LLMs are becoming the interface between natural language and decision models. Research consolidates three pathways — automatic modeling, auxiliary optimization, and direct solving — by which LLMs translate problem descriptions into mathematical models and executable code [22]; model-level systems such as ORLM convert natural-language problems into solver-compatible code and update models under constraint changes [23]; and LLMs reconfigure algorithmic components, support preference-based choices, and explain rationale for rich vehicle-routing problems [24]. In human–AI collaboration for inventory optimization under disruptions, comparing fully automated direct optimization, two-stage automated simulation and optimization, and simulation-based human-specified optimization shows SHSO consistently achieves statistically optimal, stable outcomes — evidence that retaining the human in the loop improves decision reliability [25]. Conversational planning systems built on agentic frameworks let operations planners run counterfactual reasoning and what-if/why-not scenario analysis through natural language, with human-in-the-loop generation of new tools for unsupported queries [26]. When LLMs generate optimization models directly, they exhibit a taxonomy of hallucinations with documented mitigation strategies — LLM-as-a-Judge, few-shot learning, tool calling, and multi-agent frameworks [27]; governed LLM-based optimization reduces unsafe decision outputs by 45% while sustaining drift-detection accuracy above 92% [28].

➢ *Text Mining, News Streams, and Early Warning*

Retrieval-based intelligence extends a longer line of text-mining research in supply chain risk management: systematic reviews chronicle sentiment analysis, topic modeling, and transformer-based models over Twitter and news sources for real-time risk visibility [29], deep neural networks extract supply chain maps from news articles [30], disinformation detection protects against false disruption signals [31], and news-stream deep models improve food-insecurity prediction up to 12 months ahead [32]. The portfolio's multimodal early-warning engine already reaches 92.1% recall and 93.4% precision with a 42-minute average prediction lead time [33], and neural early-warning architectures deliver two-to-four weeks of advance warning of high-impact disruptions [34]. RAG sits on top of these engines: it converts their alerts and the raw documents behind them into grounded, queryable, decision-ready intelligence.

➢ *Research Gap*

The literature separates retrieval, reasoning, and decision support (LLM–OR integration): RAG papers stop at answer quality [1], [6], agent papers stop at end-to-end analysis [2], and LLM–OR papers assume the plan rather than the evidence [22], [26]. What is missing is a retrieval-first, governed, real-time architecture that couples continuously refreshed evidence, hybrid retrieval, grounded generation, agentic orchestration, and OR-based decision support in one loop — the convergence this paper proposes, and the direction the portfolio's LLM thread has been building toward.

➢ *Problem Formulation*

- *Real-Time Disruption Intelligence As A Retrieval Problem*

Let $\mathcal{D}_t$ be the knowledge corpus at time t — news streams, supplier documents, contracts, operational records, and a knowledge graph $\mathcal{K}_t$ encoding the supply network. A user or agent issues a query q ("which of my tier-2 suppliers are exposed to the port closure announced this morning?"). The intelligence system must produce a grounded answer $\hat{y} = \mathrm{Gen}(\mathrm{prompt}(q, \mathcal{R}(q, \mathcal{D}_t)))$, where <span data-type="inline-math" data-latex="\mathcal{R} retrieves the top-k most relevant evidence items, and <span data-type="inline-math" data-latex="\mathrm{Gen} is the generation model.

- *Constraints*

Three constraints are binding in the real-time setting:

✓ Timeliness: the knowledge base must reflect the current event stream; an LLM whose knowledge ends at its training cutoff cannot answer about today's disruption [1]. The system incurs a freshness lag <span data-type="inline-math" data-latex="\tau_{\text{fresh}}(d)\le \tau_{\text{fresh}}^{\max} per document <span data-type="inline-math" data-latex="d.

✓ Faithfulness: every claim in <span data-type="inline-math" data-latex="\hat{y} must be supported by retrieved evidence; ungrounded generation is a hallucination and is mitigated by retrieval itself [7] and by knowledge-graph grounding [3], [5].

✓ Decision utility: the answer must terminate in an action. Following the LLM–OR literature, <span data-type="inline-math" data-latex="\hat{y} is transpiled into

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep297

a decision model for the planner [22], [23], or routed to a human-in-the-loop approval [25], [26].

- *Optimization Objectives*

The framework optimizes a weighted objective over retrieval, generation, and decision stages:

$$\min \mathbb{E}[\lambda_q \ell_{retr}(q) + \lambda_f \ell_{faith}(\hat{y}, \mathcal{R}) + \lambda_d (c_{delay}\tau_d + c_{err} e_d)],$$

Where <span data-type="inline-math" data-latex="\ell_{\text{retr}}} is retrieval loss (recall/mRR), <span data-type="inline-math" data-latex="\ell_{\text{faith}}} is the proportion of claims unsupported by retrieved evidence, <span data-type="inline-math" data-latex="\tau_d is decision latency, and <span data-type="inline-math" data-latex="e_d is decision error — the trade-off structure that makes retrieval quality the commercially decisive variable [1].

- *Evaluation Metrics*

✓ Retrieval: precision@k, recall@k, mRR [1], [4]. Generation: ROUGE, METEOR, and truthfulness scores [3], [5]; human-evaluation and BERTScore for extraction quality [16]. Intelligence quality: detection F1 and response accuracy [2]. Operational: end-to-end latency, cost per disruption [2], decision latency reduction [9], and retrieval time as a function of corpus size [4]. Governance: unsafe-output rate and drift-detection accuracy [28].

III. PROPOSED FRAMEWORK: RAGENT-SC

➢ *Overview*

RAGENT-SC is a retrieval-augmented agentic framework with five layers: the knowledge layer, the hybrid retrieval layer, the grounded generation layer, the agentic orchestration layer, and the governance layer. It consumes the alerts and documents of the portfolio's early-warning engines and emits grounded, actionable intelligence for the agentic decision and recovery loop and the cascade engine. Fig. 1 shows the architecture.

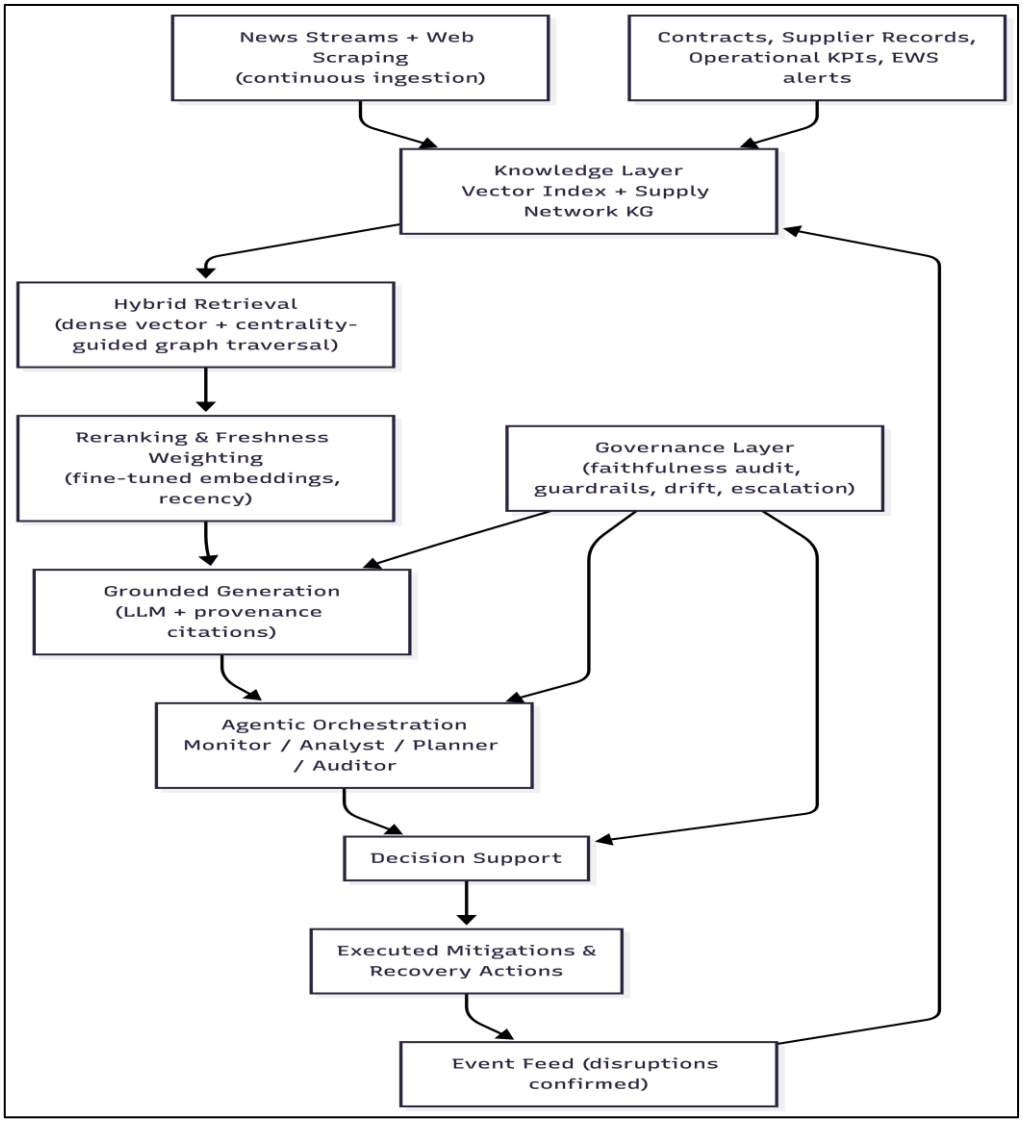

Fig 1 RAGENT-SC

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep297

- Architecture: Continuously ingested news and documents populate a knowledge layer of vector indexes and a supply-network knowledge graph; hybrid retrieval (dense vectors + centrality-guided graph traversal) feeds reranking and grounded generation with provenance; agents orchestrate monitoring, analysis, planning, and audit; decision support transpiles grounded answers into OR models with human-in-the-loop control; confirmed events feed back into the knowledge base; a governance layer audits faithfulness and enforces guardrails.

➢ *Knowledge Layer*

The layer maintains two complementary indexes: a vector index over token-chunked documents — news, contracts, supplier records, operational KPIs, and early-warning alerts — and a knowledge graph over entities and relations: companies, locations, materials, products, mines, and capabilities [14], with products, production capabilities, and certifications as in the portfolio's neurosymbolic design [35]. Because LLM knowledge is constrained to the training date, the layer is continuously updated from news streams and web sources to minimize freshness lag [1]. Ontology-driven construction, following the supplier-discovery methodology [12], ensures the graph supports queries beyond who-supplies-whom.

➢ *Hybrid Retrieval Layer*

Retrieval fuses dense vector search with graph traversal — exploiting the duality between supply networks and knowledge graphs [11]. A graph traverser guided by network-centrality scores extracts economically salient risk paths through the supplier network, while dense retrieval covers unstructured text; experiments on hybrid vector–graph structures show sublinear scaling — when the number of nodes grows fourfold, query time grows only twofold (0.09 s to 0.20 s) — with an optimal traversal budget of 30–50 nodes balancing precision and speed [4]. Retrieval quality is the system-level lever: fine-tuning the embedding retrieval model provides the most significant performance gains, and adaptive iterative retrieval adjusts retrieval depth by query complexity [1].

➢ *Grounded Generation Layer*

The generator conditions on retrieved evidence and emits every answer with provenance — source documents and graph paths — making each decision auditable, the property that enterprise-grade supplier risk assessment demands [36]. Graph grounding measurably improves generation: ROUGE-1 F1 rises from 0.42 to 0.67 (+59%) for graph-grounded LLM reasoning [3], and knowledge-graph-augmented RAG outperforms plain RAG on supply chain QA with open-weight models closing the gap with frontier models [5]. Context shells, which embed raw numeric figures in natural language, make quantitative supply chain data fully intelligible to the LLM [11].

➢ *Agentic Orchestration Layer*

Seven specialized agents, in the spirit of the validated multi-agent framework [2], coordinate the loop: a Monitor agent converts retrieved signals into candidate disruptions; an Analyst agent (with knowledge-graph and memory retrieval, following the memory-augmented agent architectures in the portfolio [20], [21]) produces structured risk assessments; a Planner agent drafts mitigations (alternative sourcing, rebalancing, re-routing); a Negotiator agent arbitrates between facility-level and systemic objectives, the consensus mechanism shown to reduce the bullwhip effect [19]; and an Auditor agent verifies faithfulness and compliance. Orchestration follows the LangGraph-style graph workflow demonstrated in LLM-powered supply chain dashboards [18], and large-corpus monitoring follows the SmartOps template — pre-filtering, NER, and LLM classification over a ~3.88M-row news repository [17].

➢ *Decision Support and Human-in-the-Loop*

Grounded intelligence terminates in decision models: the Planner's output is transpiled into solver-compatible optimization code following the LLM–OR pathways [22], [23], with counterfactual what-if/why-not analysis provided conversationally [26]. Consistent with the evidence that simulation-based human-specified optimization achieves statistically optimal outcomes [25], high-stakes actions retain human approval; the governance layer enforces the guardrails — unsafe outputs cut by 45% and drift detection above 92% in governed LLM optimization [28].

➢ *Governance layer*

The Auditor agent scores generation faithfulness against retrieved evidence, flags unsupported claims, and routes uncertain outputs to human review. Provenance metadata accompanies every risk decision to ensure auditability and regulatory compliance [36], and the ethical considerations of RAG deployment — data security and exploitation countermeasures — are treated as first-class constraints [6]. Algorithm 1 summarizes the loop.

➢ *Algorithm 1: RAGENT-SC Real-Time Intelligence Loop*

```
Input: knowledge layer (vector index V_t, graph K_t), query stream Q,
      early-warning alerts A_t, guardrails Gv
Output: grounded intelligence ŷ, executed mitigations a
Loop at decision cadence:
  # 1. INGEST
  update V_t, K_t from news/web/EWS feeds (minimize freshness lag)
  # 2. RETRIEVE
  top-k ← dense(V_t, q) ⊕ centrality-traversal(K_t, q)  # budget 30–50 nodes
  rerank ← fine-tuned embedding model + recency weighting
  # 3. GENERATE
  ŷ ← Gen(prompt(q, reranked evidence)); attach provenance (doc, path)
  # 4. VERIFY
  faithfulness ← Auditor(ŷ, evidence)
  if faithfulness < τ: escalate to human review
  # 5. ORCHESTRATE
  risk_assessment ← Analyst(ŷ, K_t, memory)
  plan ← Planner(risk_assessment)            # alternative sourcing, rebalancing
  if not Guardrails.ok(plan): plan ← HITL(plan)
```

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep297

```
# 6. DECIDE & ACT
a ← transpile(plan) → OR solver (Gurobi/COPT) or DRL policy
execute a; feed confirmed events back to V_t, K_t
# 7. LEARN
fine-tune embeddings periodically; update drift detectors
```

➢ *Implementation*

Embeddings: fine-tuned retrieval model [1] with FAISS/Llama Index indexing [18]; graph: property graph with centrality-based traversal [11] and a 30–50-node budget [4]; generator: LLM with top-p decoding and provenance templates [11], [36]; agents: LangGraph-style orchestration [18]; decision layer: LLM-to-OR transpilation [22], [23]; governance: faithfulness auditor, guardrail checker, drift detector [28], [36].

➢ *Experimental Setup*

- *Environments and Data*

Evaluation spans the documented settings of the component literature: the seven-agent disruption-monitoring evaluation across 30 synthesized scenarios covering three automotive manufacturers and five disruption classes [2]; the 1,277-sentence knowledge-graph extraction dataset for electric-vehicle supply chains [14]; the SCM benchmark of eight core supply chain functions and the LTU chatbot QA dataset [5]; the longitudinal Indian news repository of ~3.88 million rows (2001–2023) [17]; the natural-hazard news-retrieval setting with human and BERTScore evaluation [16]; and the retrieval-scaling testbed of the hybrid vector–graph study [4].

- *Baselines*

Plain (non-retrieval) LLM prompting and training-date-constrained LLMs [1]; dense-only retrieval and graph-only retrieval versus hybrid [4], [11]; plain RAG versus knowledge-graph-augmented RAG [3], [5]; single-prompt versus chain-of-thought and few-shot extraction [16]; and rule-based/text-mining early warning as the intelligence baseline [29], [33].

- *Protocol*

Temporal splits prevent leakage [2]; retrieval is evaluated on held-out queries with precision@k and mRR [1], [4]; generation is scored with ROUGE/METEOR and truthfulness metrics [3], [5]; extraction is scored by human evaluation and BERTScore [16]; end-to-end latency and cost are measured per disruption [2]. All numerical entries below are literature-reported values from the cited studies.

IV. RESULTS AND ANALYSIS

➢ *Task–Module Mapping*

Table 1 maps each framework module to its task and evaluation.

Table 1 RAGENT-SC Modules, Tasks and Evaluation

Module	Task	Output	Evaluation
Knowledge layer	Multi-format corpus + KG updates	V_t, K_t	Freshness lag, KG coverage
Hybrid retrieval	Evidence retrieval	top-k with provenance	precision@k, mRR, scaling
Grounded generation	Factual synthesis	$\hat{y}$ with citations	ROUGE/METEOR, truthfulness, faithfulness
Agentic orchestration	Monitoring/analysis/planning	risk assessments, plans	Detection F1, response accuracy
Decision support	Plan → OR model	executed mitigations	Decision latency, error rate
Governance	Faithfulness audit, guardrails	approval/escalation	Unsafe-output rate, drift accuracy

➢ *Reported Performance Envelope*

Table 2 consolidates reported results across the RAG/LLM supply chain literature.

Table 2 Reported Performance in Retrieval-Augmented Supply Chain Intelligence

Method	Task	Reported result	Source
Seven-agent LLM framework	Multi-tier disruption monitoring & response	F1 0.962–0.991; 3.83 min; $0.0836 per disruption; >$10^3$ faster than multi-day analyst	[2]
Graph-based LLM reasoning	Entity classification, link prediction, reasoning	ROUGE-1 F1 0.42 → 0.67 (+59%)	[3]
Hybrid vector–graph retrieval	NL information search in logistics	4× nodes → 2× time (0.09 → 0.20 s); 30–50-node optimum; best accuracy among methods	[4]
Real-time RAG + web scraping	Timely vulnerability identification	Embedding-retriever fine-tuning = most significant gains; adaptive iterative retrieval helps complex queries	[1]
SC-KAE (KG + RAG)	SCM QA (8 core functions + LTU)	KG integration outperforms plain RAG; smaller models close gap with larger ones	[5]

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep297

Domain-specialized SCM LLM	SCM exams + beer game	Passes standardized examinations; RAG significantly improves SCM task performance	[7]
RAG + KG decision support	Cross-domain decision support	Better accuracy, transparency, relevance vs either alone	[13]
Natural-hazard LLM extraction	Risk-feature extraction from news	Llama3.1-8B & Phi4-14B top human eval + BERTScore; DeepSeek-R1-14B lowest	[16]
SmartOps Risk Monitor	News-driven risk awareness	~3.88M rows (2001–2023); regulatory/ecological risks primary; ESG 61.9/100; energy R^2 0.797	[17]
SupChain-Bench / ReAct	SOP-grounded tool orchestration	Substantial gaps in execution reliability; SOP-free ReAct strongest tool-calling	[8]
SMARTAPS	Conversational APS	NL query, counterfactual, what-if/why-not; HITL tool generation	[26]
Eval-RCoT	Supplier qualification (LLM + CoT + RAG)	Accurate rating levels + interpretable insights	[37]
RAG for SC operations	Forecasting, inventory, supplier risk	Dual-component RAG; iterative retrieval; chunk optimization	[6]
Hybrid RAG-LLM supplier risk	Geopolitical/financial/compliance risk	Provenance metadata per decision; structured governance	[36]
Network–KG agentic analysis	Real-time risk narratives	Centrality-guided traversal; context shells; no fine-tuning	[11]

➢ *Worked Calculations*

- Response-time reduction (the headline number). An end-to-end analysis in a mean of 3.83 minutes constitutes a reduction of more than three orders of magnitude relative to multi-day analyst-driven assessments [2]. Working backward, the implied benchmark for a comparable analyst assessment is at least $$3.83\ \text{min} \times 10^{3} = 3{,}830\ \text{min} \approx 63.8\ \text{hours} \approx 2.7\ \text{days},$$ consistent with the stated multi-day industry benchmark, and with the industry-reported shift from days to minutes for LLM-based decision support [9].
- Grounding gain arithmetic. Graph-grounded generation raises ROUGE-1 F1 from 0.42 to 0.67 [3]. The absolute gain is +0.25; the relative gain is $(0.67 - 0.42)/0.42 \approx 59.5\%$, matching the reported 59% improvement over the average of the best comparative methods. In retrieval terms, this is what grounding is worth: roughly a 60% relative lift in answer fidelity from providing structured, factual context to the generator — consistent with the finding that knowledge-graph augmentation improves plain RAG across SCM question answering [5].
- Retrieval scalability arithmetic. The hybrid vector–graph method doubles query time from 0.09 s to 0.20 s when node count quadruples [4]. The implied scaling law is $t \propto N^{\alpha}$ with i.e., sublinear growth: the graph-traversal budget (30–50 nodes) bounds the search region so that corpus growth does not linearly tax query latency [4]. At live-news volume, this is the property that keeps retrieval inside the real-time decision cadence.

$$\alpha = \frac{ln(0.20/0.09)}{ln\,4} \approx \frac{0.799}{1.386} \approx 0.58,$$

- Unit economics. At a mean cost of $0.0836 per disruption [2], sustained monitoring across even thousands of candidate events per year remains in the tens to hundreds of dollars — orders of magnitude below a single analyst-day. Retrieval-grounded automation therefore makes continuous disruption intelligence economically feasible, not just technically possible.
- The retrieval-quality lever. Fine-tuning the embedding retrieval model delivers the most significant performance gains in RAG supply chain analysis, and adaptive iterative retrieval further improves complex queries [1]. Because detection F1 already sits at 0.962–0.991 [2], incremental accuracy gains now come from the evidence layer: better retrieval, fresher corpora, and denser knowledge graphs — the argument that retrieval quality, not model scale, is the binding constraint.
- Scaling the monitoring corpus. The SmartOps template processes a ~3.88M-row news repository (2001–2023) with pre-filtering, NER, and LLM classification [17]. At national-news scale, retrieval-grounded pipelines make continuous risk classification tractable; the RAGENT-SC Monitor agent adopts the same three-stage pipeline with the graph-traversal budget of Section IV-C to bound per-query cost.

➢ *Grounding Gain*

Fig. 2 plots the reported ROUGE-1 improvement from graph grounding.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep297

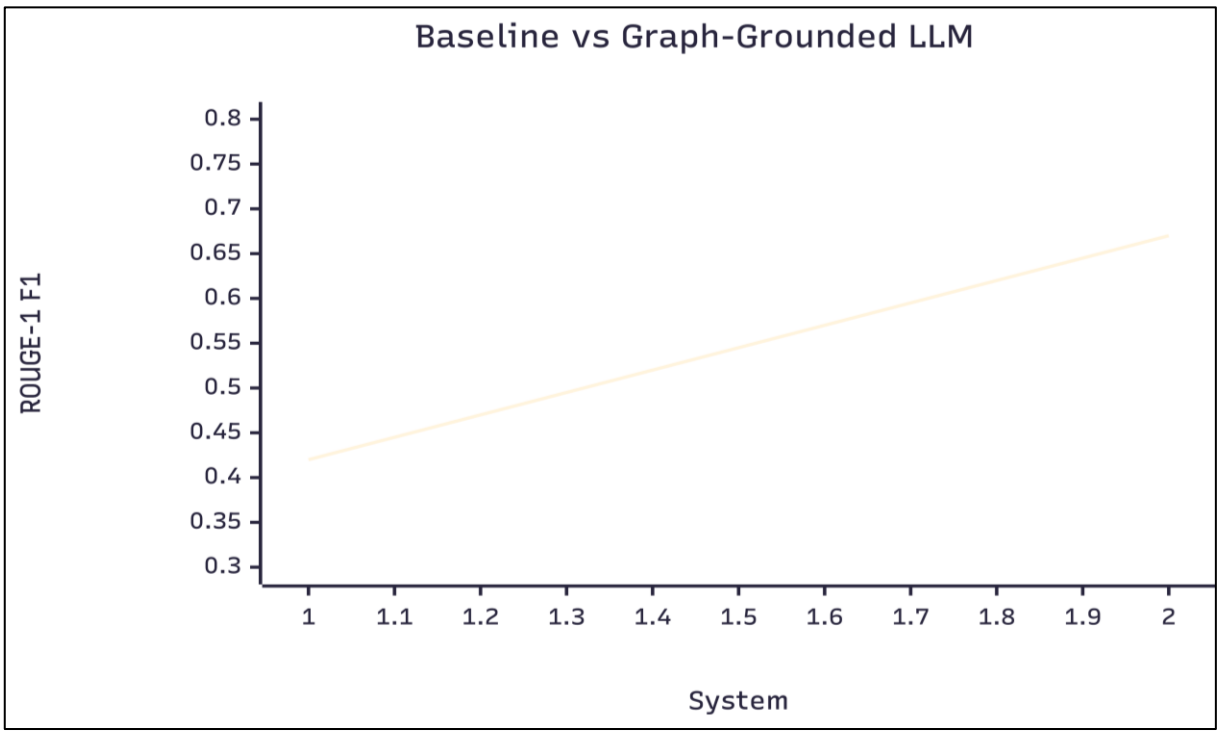

Fig 2 Reported ROUGE-1 F1 for Graph-Grounded LLM Reasoning vs the Average of the Best Comparative Methods: 0.42 → 0.67 (+59%) [3].

➢ *Retrieval Scaling*

Fig. 3 shows the reported sublinear scaling of the hybrid vector–graph retrieval method.

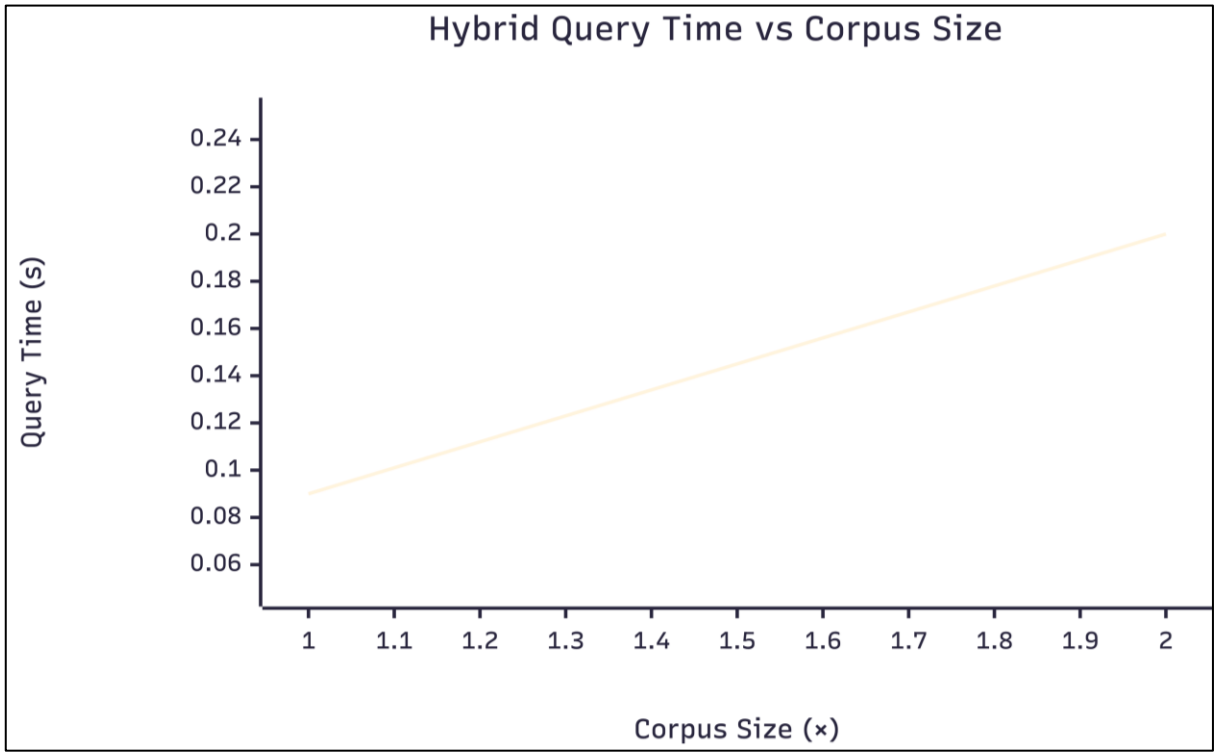

Fig 3 Reported Query Time of the Hybrid Vector–Graph Method: 0.09 s at Baseline to 0.20 s at 4× Node Count — Sublinear Scaling ($\alpha \approx 0.58$) [4].

➢ *Illustrative Intelligence Latency*

Fig. 4 shows the [illustrative] cumulative intelligence accuracy over time for an event, contrasting analyst-driven assessment (multi-day) with RAGENT-SC (retrieval-grounded, agent-minutes), consistent with the reported three-orders-of-magnitude latency reduction [2].

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep297

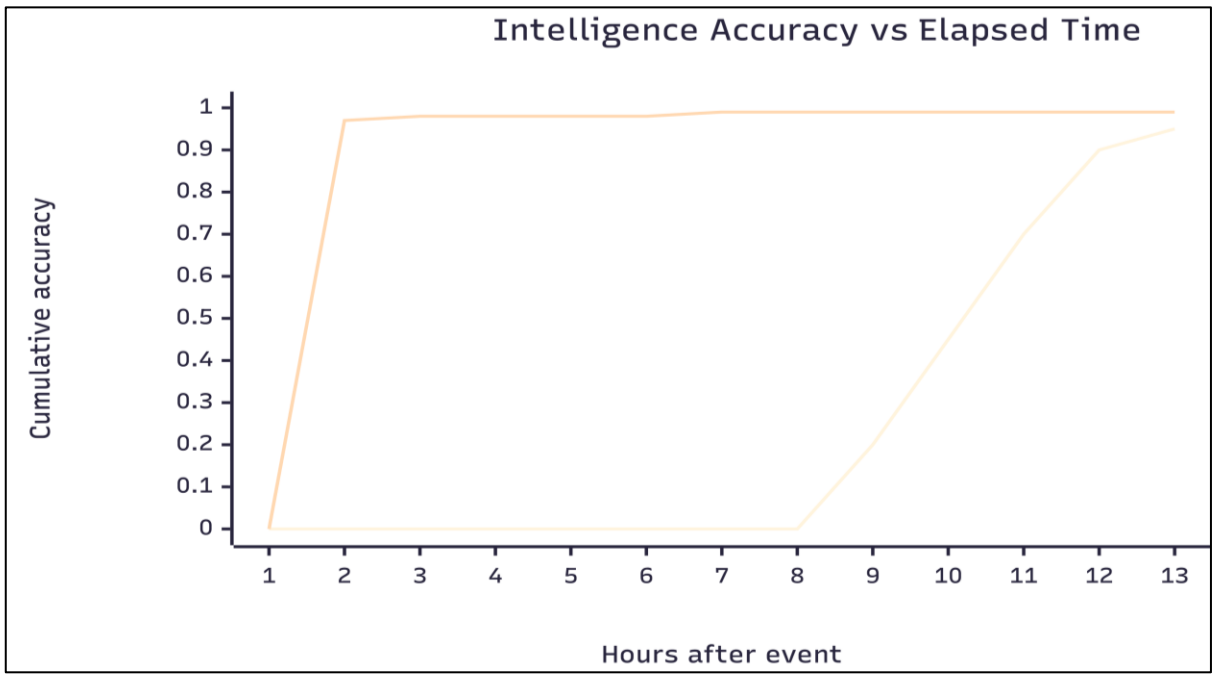

Fig 4 [Illustrative] Cumulative Disruption-Intelligence Accuracy Following an Event: Analyst-Driven Assessment Reaching High Fidelity Only after ~2.7 Days (First Line) Versus Retrieval-Grounded Agentic Analysis Reaching Reported F1 Fidelity (0.962–0.991) within Minutes (Second Line) [2]

➢ *Ablation Logic*

Removing the retrieval layer collapses the system to a training-date-constrained LLM that cannot reason about current events [1]; removing the graph index degrades relational queries to vector-distance approximations [11], consistent with the ROUGE-1 gain attributable to graph grounding [3]; removing fine-tuned embeddings forfeits the most significant retrieval-quality gains [1]; and removing the governance layer reintroduces the hallucination and unsafe-output risks that retrieval grounding and guardrails are designed to contain [7], [28], [36].

V. DISCUSSION

➢ *Retrieval Quality is the Binding Constraint*

The evidence converges on one conclusion: at F1 0.962–0.991 for agentic detection [2] and ROUGE-1 0.67 for grounded generation [3], the marginal value of a bigger model is smaller than the marginal value of a better retriever — fine-tuning the embedding model is the most significant performance lever in RAG supply chain analysis [1], and knowledge-graph grounding lifts plain RAG across SCM benchmarks [5]. The research program should therefore invest in the evidence layer: corpus freshness, graph density, centrality-guided traversal, and context-shell encoding of quantitative data [11].

➢ *Real-Time Requires Continuous Ingestion*

An LLM's parametric knowledge is frozen [1]; a disruption-intelligence system's knowledge cannot be. The portfolio's early-warning engines detect signals with quantified lead time [33], [34]; RAGENT-SC supplies the why in real time by retrieving the documents behind those signals. The news-driven monitoring template — pre-filtering, NER, LLM classification over millions of rows [17] — and the natural-hazard extraction results showing prompt-strategy and model-choice sensitivity [16] define the ingestion engineering that makes retrieval-grounded intelligence trustworthy.

➢ *Grounding is the Governance Mechanism*

- *Retrieval Does Double Duty:*

It improves accuracy and it creates auditability. Provenance metadata that accompanies every risk decision [36], faithfulness scoring that flags unsupported claims, and guardrail-based escalation that retains human control [25], [28] are what move RAG from a research artifact to a deployable decision-support system. The hallucination taxonomies and mitigation strategies of the LLM-optimization literature — LLM-as-a-Judge, few-shot grounding, tool calling, multi-agent verification [27] — apply directly to the Planner agent.

➢ *The Human Stays in the Loop — Selectively*

The comparison of fully automated, two-stage, and human-specified LLM-based inventory optimization shows that human-specified algorithms achieve statistically optimal and stable outcomes under disruption [25]. RAGENT-SC encodes this selectively: routine monitoring is autonomous, but high-stakes mitigation plans route through the Auditor and human approval, and the conversational planning interface [26] gives planners counterfactual and what-if/why-not control without OR-consultant dependency [9], [26].

➢ *Integration with the Portfolio*

RAGENT-SC is the evidence and reasoning spine of the seven-paper program: it grounds Paper 1's alerts in retrievable documents; supplies the retrieval and memory substrate for Paper 2's agentic decision loop [20], [21]; feeds Paper 3's

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep297

cascade engine with freshly extracted supply-network relations (mining the same unstructured sources that link prediction reconstructs [38], [39]); enriches Paper 4's weather front end with news-derived hazard confirmation [16], [31]; and extends Paper 5's news-stream food-crisis intelligence [32] to the general commodity domain.

➢ *Limitations*

Reported results span heterogeneous tasks, datasets, and models rather than a single controlled benchmark; RAG's application to supply chains is limited by the absence of high-quality, domain-specific knowledge bases [7]; benchmark work on long-horizon tool orchestration exposes substantial execution-reliability gaps [8]; and multi-LLM comparisons show model choice materially affects extraction quality [16]. These are the open problems of Section VIII.

VI. CONCLUSION AND FUTURE WORK

This paper formalized real-time supply chain disruption intelligence as a retrieval-grounded generation problem and proposed RAGENT-SC, a retrieval-augmented agentic framework coupling continuously updated knowledge, hybrid vector–graph retrieval, grounded generation with provenance, agentic orchestration, decision support, and governance. It consolidated the reported evidence envelope — F1 0.962–0.991 with end-to-end analysis in 3.83 minutes at $0.0836 per disruption and a greater-than-10^3 latency reduction [2]; ROUGE-1 0.42 → 0.67 from graph grounding [3]; sublinear hybrid-retrieval scaling (0.09 → 0.20 s at 4× nodes) with a 30–50-node traversal budget [4]; retrieval-model fine-tuning as the most significant RAG performance lever [1]; KG-augmented RAG outperforming plain RAG with smaller models closing the frontier gap [5]; and domain-specialized SCM LLMs passing standardized examinations [7] — and argued that retrieval quality, not model scale, is the binding constraint on real-time disruption intelligence.

Future work will (i) implement RAGENT-SC end-to-end on a live commodity supply chain with unified evaluation across the Bench-style protocols of Section V [8]; (ii) build and continuously refresh high-quality domain knowledge bases — the documented bottleneck of supply chain RAG [7]; (iii) deepen graph retrieval with centrality-guided and uncertain-knowledge-graph traversal [11], [39]; (iv) couple retrieval grounding with the portfolio's LLM-to-OR decision support so grounded plans transpile directly into solver-verified actions under guardrails [22], [23], [28]; and (v) standardize evaluation — detection F1, faithfulness, latency, cost, and decision error — so that retrieval-augmented disruption intelligence becomes a measurable, auditable, and deployable science, converting LLMs from recall machines into real-time resilience engines.

REFERENCES

[1]. J. Ponnock et al. , “Real-Time RAG for the Identification of Supply Chain Vulnerabilities,” Aug. 23, 2025. doi: 10.48550/arxiv.2509.10469 .

[2]. S. AlMahri, L. Xu, and A. Brintrup, “Automating Supply Chain Disruption Monitoring via an Agentic AI Approach,” ArXiv.org , Jan. 2026, Accessed: Feb. 2026. [Online]. Available: http://arxiv.org/abs/2601.09680

[3]. P. P. Su, R. Xu, and D. Chen, “Leveraging the Graph-Based LLM to Support the Analysis of Supply Chain Information,” Informatics , vol. 12, no. 4, p. 124, Nov. 2025, doi: 10.3390/informatics12040124 .

[4]. V. I. Voloshchuk, Y. E. Melnik, I. Safronenkova, E. Lishchenko, O. O. Kartashov, and A. Kozlovskiy, “Hybrid Method of Organizing Information Search in Logistics Systems Based on Vector-Graph Structure and Large Language Models,” Big Data and Cognitive Computing , vol. 10, no. 2, p. 51, Feb. 2026, doi: 10.3390/bdcc10020051 .

[5]. M. Sushanta, “Augmenting Large Language Models with Domain-Specific Insight: Establishing SC-KAE Framework for Improved Real-World Application of LLMs in Supply Chain,” KTH Publication Database DiVA (KTH Royal Institute of Technology) , Jan. 2025, Accessed: Mar. 2026. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:ltu:diva-115277

[6]. B. Zhu and C. Vuppalapati, “Enhancing Supply Chain Efficiency Through Retrieve-Augmented Generation Approach in Large Language Models,” pp. 117–121, Jul. 2024, doi: 10.1109/bigdataservice62917.2024.00025 .

[7]. H. Wang, J. Jiang, L. J. Hong, and G. Jiang, “LLMs for Supply Chain Management,” May 24, 2025. doi: 10.48550/arxiv.2505.18597.

[8]. S. Guan, Y. Liu, and L. Cao, “SupChain-Bench: Benchmarking Large Language Models for Real-World Supply Chain Management,” arXiv (Cornell University) , Feb. 2026, Accessed: Feb. 2026. [Online]. Available: http://arxiv.org/abs/2602.07342

[9]. T. Maiti, “Application of Large Language Models (LLM’s) for Supply Chain Optimization,” INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT , vol. 9, no. 5, pp. 1–9, May 2025, doi: 10.55041/ijsrem47807 .

[10]. L. Xu, S. AlMahri, S. Mak, and A. Brintrup, “Multi-Agent Systems and Foundation Models Enable Autonomous Supply Chains: Opportunities and Challenges,” IFAC-PapersOnLine , vol. 58, no. 19, pp. 795–800, Jan. 2024, doi: 10.1016/j.ifacol.2024.09.200 .

[11]. E. Heus, R. Bookstaber, and D. Sharma, “Exploring Network-Knowledge Graph Duality: A Case Study in Agentic Supply Chain Risk Analysis,” Oct. 01, 2025. doi: 10.48550/arxiv.2510.01115 .

[12]. Y. Li, H. Ko, and F. Ameri, “Integrating Graph Retrieval-Augmented Generation With Large Language Models for Supplier Discovery,” Journal of Computing and Information Science in Engineering , vol. 25, no. 2, Dec. 2024, doi: 10.1115/1.4067389 .

[13]. S. Wang, H. Yang, and G. Bai, “Construction of intelligent decision support systems through integration of retrieval-augmented generation and knowledge graphs,” Scientific Reports , vol. 15, no. 1,

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep297

p. 35462, Oct. 2025, doi: 10.1038/s41598-025-19257-3 .

[14]. S. AlMahri, L. Xu, and A. Brintrup, "Enhancing Supply Chain Visibility with Knowledge Graphs and Large Language Models," Aug. 05, 2024, Cornell University . doi: 10.48550/arxiv.2408.07705 .

[15]. Z. Ge and A. Brintrup, "Enhancing Supply Chain Visibility with Generative AI: An Exploratory Case Study on Relationship Prediction in Knowledge Graphs," Dec. 04, 2024, Cornell University . doi: 10.48550/arxiv.2412.03390 .

[16]. E. E. Günay et al. , "Enhancing Supply Chain Risk Management with LLMs: An Application for Natural Hazard Monitoring," Journal of Computing and Information Science in Engineering , pp. 1–65, Feb. 2026, doi: 10.1115/1.4071183 .

[17]. K. S. Rajesh, S. Surve, J. Joy, and A. Munshi, "SmartOps: From Demand to Delivery," pp. 1–8, Dec. 2025, doi: 10.1109/icaiqsa67794.2025.11440525 .

[18]. J. N. T. U. Hyderabad, "LLM-Powered Supply Chain Dashboard: A Graph-based multi-module framework for Data-driven Decision making," Nov. 24, 2025, European Organization for Nuclear Research . doi: 10.5281/zenodo.17695657 .

[19]. V. Jannelli, S. Schöpf, M. Bickel, T. H. Netland, and A. Brintrup, "Agentic LLMs in the supply chain: towards autonomous multi-agent consensus-seeking," International Journal of Production Research , pp. 1–31, Dec. 2025, doi: 10.1080/00207543.2025.2604311

[20]. K. Yoshizato, K. Shimizu, R. Higa, and T. Otsuka, "AI Agent Systems for Supply Chains: Structured Decision Prompts and Memory Retrieval," Feb. 05, 2026, Cornell University . doi: 10.48550/arxiv.2602.05524 .

[21]. N. K. Jingar, "Automated Incident Intelligence In Supply Chains Using Agentic AI And Root Cause Reasoning," Zenodo (CERN European Organization for Nuclear Research) , Sep. 2023, doi: 10.5281/zenodo.18628070 .

[22]. Y. Wang and K. Li, "Large Language Models in Operations Research: Methods, Applications, and Challenges," Sep. 18, 2025, Cornell University . doi: 10.48550/arxiv.2509.18180 .

[23]. Y. Yan et al. , "Large Language Models for Traffic and Transportation Research: Methodologies, State of the Art, and Future Opportunities," Mar. 27, 2025. doi: 10.48550/arxiv.2503.21330 .

[24]. G. Ghiani, E. Manni, and S. Zacchino, "Improving Adaptability in Optimization-Based Decision Support Systems Through Large Language Models," IEEE Access , vol. 13, pp. 149777–149788, Jan. 2025, doi: 10.1109/access.2025.3601518 .

[25]. R. Wu, "Inventory optimization under supply chain disruptions: Leveraging large language models for human-AI collaborative decision-making," Journal of King Saud University - Computer and Information Sciences , vol. 38, no. 2, Jan. 2026, doi: 10.1007/s44443-025-00458-9 .

[26]. T. T. Yu et al. , "SMARTAPS: Tool-augmented LLMs for Operations Management," Jul. 23, 2025. doi: 10.48550/arxiv.2507.17927 .

[27]. G. Zhang, "Large Language Model enabled Mathematical Modeling," Oct. 22, 2025. doi: 10.48550/arxiv.2510.19895 .

[28]. N. K. Jingar, "Ensuring Safety, Accountability, and Drift Resistance in LLM-Based Supply Chain Optimization," International Journal of Scientific Research in Science Engineering and Technology , p. 472, Jan. 2023, doi: 10.32628/ijsrset2310372 .

[29]. G. Gelastopoulos and C. Keramydas, "A systematic review of text mining analytics for supply chain risk management using online data," Supply Chain Analytics , vol. 12, p. 100167, Sep. 2025, doi: 10.1016/j.sca.2025.100167 .

[30]. P. Wichmann, A. Brintrup, S. Baker, P. Woodall, and D. McFarlane, "Extracting supply chain maps from news articles using deep neural networks," International Journal of Production Research , vol. 58, no. 17, pp. 5320–5336, Feb. 2020, doi: 10.1080/00207543.2020.1720925 .

[31]. P. Akhtar et al. , "Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions," Annals of Operations Research , vol. 327, no. 2, pp. 633–657, Nov. 2022, doi: 10.1007/s10479-022-05015-5 .

[32]. A. Balashankar, L. Subramanian, and S. P. Fraiberger, "Predicting food crises using news streams," Science Advances , vol. 9, no. 9, Mar. 2023, doi: 10.1126/sciadv.abm3449 .

[33]. J. Wang, "Multimodal Deep Learning Approach for Early Warning of Supply Chain Disruptions Using NLP and Anomaly Detection," Artificial Intelligence and Machine Learning Review , vol. 5, no. 3, pp. 98–110, Jul. 2024, doi: 10.69987/aimlr.2024.50308 .

[34]. Sichong Huang, "AI-Driven Early Warning Systems for Supply Chain Risk Detection: A Machine Learning Approach," Academic Journal of Computing & Information Science , vol. 8, no. 9, Jan. 2025, doi: 10.25236/ajcis.2025.080913 .

[35]. E. E. Kosasih, F. Margaroli, S. Gelli, A. Aziz, N. Wildgoose, and A. Brintrup, "Towards knowledge graph reasoning for supply chain risk management using graph neural networks," International Journal of Production Research , vol. 62, no. 15, pp. 5596–5612, Jul. 2022, doi: 10.1080/00207543.2022.2100841 .

[36]. S. Vadakkepati, "Hybrid RAG-LLM Framework For Intelligent Supplier Risk Assessment In Global Supply Chains," Journal of International Crisis and Risk Communication Research , pp. 72–87, Dec. 2025, doi: 10.63278/jicrcr.vi.3496 .

[37]. L. Xuerong, S. Wei, N. Zihang, X. Bingxin, and Z. Lexiang, "Can large language models evaluate suppliers' qualifications of complex production lines? An approach incorporating retrieval-augmented generation and Chain-of-Thought," Journal of the Association for Information Systems , Dec. 2025, Accessed: Mar. 2026. [Online]. Available: https://aisel.aisnet.org/icis2025/da_bus/da_bus/5

[38]. E. E. Kosasih and A. Brintrup, "A machine learning approach for predicting hidden links in supply chain with graph neural networks," International Journal of

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep297

Production Research , vol. 60, no. 17, pp. 5380–5393, Jul. 2021, doi: 10.1080/00207543.2021.1956697 .

[39]. N. Brockmann, E. E. Kosasih, and A. Brintrup, "Supply Chain Link Prediction on Uncertain Knowledge Graph," ACM SIGKDD Explorations Newsletter , vol. 24, no. 2, pp. 124–130, Nov. 2022, doi: 10.1145/3575637.3575655 .